Reclassification of serine / threonine phosphorylation sites with +1 proline (S/T-P) sites as a distinct eukaryotic post-translational modification class

by

Min-Hyung Cho

A dissertation submitted to Johns Hopkins University in conformity with the requirement for the degree of Doctor of Philosophy

Baltimore, Maryland

March 2020

Abstract

+1 proline is the most frequently found sequence motif around serine/threonine phosphorylation sites. While these proline-directed serine/threonine (S/T-P) phosphorylation sites accounts for about 1/3 of known human phosphorylation sites, it is less frequently studied than other types of phosphorylation sites: partly because of its unclear sequence consensus and reduced probability of generating attestable phenotypes when modified.

In this study, we propose to establish this S/T-P phosphorylation sites as a distinctive subclass of protein phosphorylation by its own. We investigated sequence preferences, biophysical fingerprints & ontological associations of known human phosphorylation sites and found there is a significant difference between S/T-P sites and other serine / threonine phosphorylation sites, which would lead to difference consequences after phosphorylation.

Also, we found 'horizontal' – sequence averaged – information plays a major role in distinguishing S/T-P sites from non-phosphorylated counterparts, while other serine/threonine phosphorylation sites and tyrosine phosphorylation sites strongly rely on 'vertical' – sequence specific – information to differentiate those from non-phosphorylated counterparts. These behaviors were specifically associated with +1 proline: using proline residues on other locations or other residues on +1 site as criteria were not able to reproduce these pre-stated differences.

Furthermore, we identified not only +1 proline is evolutionarily conserved across phosphoprotein orthologs, but also S/T-P sites were slowly enriched within mammalian level. Interestingly, +1 proline is more likely to be in the reconstructed ancestral sequences than actually phosphorylated serine/threonine residues, which might imply about the possible origin and evolutionary advantage of S/T-P phosphorylation.

These results would not only provide an insight about this 'neglected subclass' of phosphorylation sites, but would also suggest this particular PTM is co-evolved with eukaryotic proteome to carry out roles associated with biological complexity.

Advisor: Prof. Vincent Hilser & Prof. James Taylor

Reader: Prof. Bertrand Garcia-Moreno & Prof. Margaret Johnson

Acknowledgements

I would like to properly express my gratitude to those who helped me during my years in graduate school. This work would not have been possible without their assistance and commitment.

First and foremost, I would like to thank my advisors, Dr. Vincent Hilser and Dr. James Taylor, for every invaluable advices, insights, opportunities and resources they provided for my transformation into a full-fledged researcher. The perspectives I have learned from them enabled me to embrace a variety of ideas from different fields of biology and consequently develop my independent thinking skill, which I consider as an essential virtue for my career. Furthermore, the environments they provided me were crucial for cultivating my own idea and discussing it freely. It has been a true privilage to work with them.

I am grateful to my thesis committee members, Dr. Bertrand Garcia-Moreno and Dr. Margaret Johnson, for providing me such effective suggestions and guidance in every thesis reviews. Their mentorship augmented my advisors' guidance, which lead me to broaden my knowledge and overcome numerous obstacles I faced. I also appreciate faculty and staff members who provided assistance in more than one way - my GBO committee members Dr. Doug Barrick, Dr. Christian Kaiser and Dr. Elijah Roberts; Dr. Andrew Gordus, Dr. Samar Hattar and Dr. Bob Johnston for thoughtful comments and questions; and Ms. Joan Miller and Ms. Barbara Birsit who have spent substantial time to aid me solving numerous issues.

Every members in my labs, including both present and previous - Dr. Enis Afgan, Dr. Jeremy Anderson, Dr. Alex Chin, Dr. Mallory Freeberg, Dr. Andrew Martens, Dr. Hesam Motlagh, Dr. Harry Saavedra, Dr. Mike Sauria, Dr. Jordan white, Sarah Brantley, Boris Brenerman, Peter Deford, Emily Grasso, Joseph Rehfus, James Rives, Nathan Roach, Miranda Russo, German Uritskiy, Keila Voortman-sheetz and Kate Weaver - deserve my gratitude. I am fortunate to have such great colleagues who let me share the ideas and keep moving forward.

Among these colleagues, I am particularly indebted to Dr. James Wrabl, who spend several years alongside me to realize my ideas to tangible results. He has been always prepared not only to share research experience

but also to discuss about results and provide invaluable insights. Our collaboration has been indeed integral to develop my research into a level I would never reach by myself.

Finally, I would like to acknowledge everyone in my personal life who have been supportive of me all the time, including my extended network of family and friends. Most importantly, not just during the years in graduate school but throughout my life, I received unfaltering encouragement and love from my parents, Byung Chul Cho and Hyung Sook Lee, and my sister, Chae Won Cho. They have been always there for me when I needed them the most.

For each person I mentioned above, and countless others I have omitted here, I dedicate my thesis to thee.

Table of contents

List of Tables

List of Figures

Abbreviations

Arg / R Arginine

ASA Accessible surface area

Asp / D Aspartate

AUROC Area under ROC curve

Glu / E Glutamate

IDR Intrinsically disordered region of protein

Lys / K Lysine

MCC Matthews correlation coefficient

nS Non-phosphorylated serine

nT Non-phosphorylated threonine

nY Non-phosphorylated tyrosine

PII Polyproline II (helix) conformation

Pro / P Proline

pS Phosphoserine

pT Phosphothreonine

PTM Post-translational modification

pY Phosphotyrosine

ROC Receiver operating characteristics

Ser / S Serine

S-nP Serine phosphorylation sites without +1 proline

S-P Serine phosphorylation sites with +1 proline

S/T-nP Serine / threonine phosphorylation sites without +1 proline

S/T-P Serine / threonine phosphorylation sites without +1 proline

Thr / T Threonine

T-nP Threonine phosphorylation sites without +1 proline

T-P Threonine phosphorylation sites with +1 proline

Tyr / Y Tyrosine

Xaa Any amino acid

[1] Backgrounds

[1-1] Post-translational modification

Proteins, which make up the majority of enzymes found in life, are involved in virtually every biological process (1). This is partly due to structural diversity and functional versatility inherent to protein: which are ultimately conferred by diversity of structural unit itself - the amino acids. Amino acids consist of amine group and carboxyl group, which are required to form peptide bond between two amino acids and thereby necessary for polymerization, and side chains (residues) specific to each amino acids, which have highly divergent physical and chemical properties. Exponential possibilities of amino acid combinations allow proteins to adopt a multitude of different three dimensional structures, and consequently, to possess variety of functions (2).

According to the central dogma proposed by Francis Crick, protein sequence information is stored in protein-coding genes in the genome, which is transcribed by RNA polymerases and subsequently translated into actual polypeptide by ribosome (3). During this translation process, ribosome could include 22 types of proteinogenic amino acids (4), which include 20 amino acids encoded by universal genetic code and 2 special amino acids – selenocysteine (Sec, U) and pyrrolysine (Pyl, O) – which are inserted into nascent polypeptide when there are specific RNA elements in the mRNA (5, 6). The problem is, the diversity of amino acids found in protein far exceeds that of proteinogenic amino acids. Mass spectrometry has identified hundreds of different amino acids in natural proteins so far, which are obviously not random inclusions during protein translation process or experimental artifacts (7).

These non-proteinogenic amino acids are mostly introduced into proteins via a mechanism called protein post-translational modification (PTM), a covalent modification of amino acid residues in translated proteins (8). PTM may occur on most of amino acid residues: the exceptions are mostly aliphatic amino acids, namely alanine, valine, leucine, isoleucine and phenylalanine (9). Also, each amino acid type could be modified in more than one way, which often results in drastically different end products. Depending on the types, PTM could be either stable, which may last indefinitely until the protein is degraded, or transient, which could be

dynamically applied and removed in response to the environmental cues (10). For this reason, PTM could be involved in many different biological processes, including gene expression, molecular translocation, signal transduction and structural organization (11). Among all these PTM types, about a dozen are being rigorously studied, which are shown in Table 1.

PTM alters physicochemical properties of target amino acids, including size, surface area, charge and hydrophobicity, which may result in changes of properties of substrate protein, such as tertiary structure, enzymatic activity, molecular lifetime and protein-protein interaction patterns (12). Examples include phosphorylation, which often function as an on/off switch of protein activity (13), disulfide bond, which forms between cysteine residues and provide structural stability (14), and ubiquitination, which regulates not only protein degradation but also other processes such as transcription, signaling and autophagy (15). These changes may consequently prompt effects of much larger scales, such as regulation of chromatin states by acetylation / methylation of histone tails (16). In some cases, multiple PTMs are functionally associated with each other and produce emergent effects, such as tau protein aggregation associated with hyperphosphorylation (17).

Along with alternative splicing, PTM is a pivotal mechanism which increases proteome complexity of eukaryote (7). While there is a disagreement in the exact number, the total number of distinct PTM sites in the human proteome is at least in the order of hundreds of thousands, if not millions (18). This suggests multiple different PTM sites would exist in each protein species, which could further lead to a combinatorial expansion of the molecular states (19). This allows millions of structurally and functionally different protein species to be derived from much smaller pool of protein coding genes, which the size is about 20,000 (20).

[1-2] Protein phosphorylation

Protein phosphorylation is the addition of phosphate group to nucleophilic amino acid residues (21) such as serine (Ser, S), threonine (Thr, T), tyrosine (Tyr, Y), aspartate (Asp, D), glutamate (Glu, E), histidine (His,

PTM types	Modified residue	Chemical donor	Intracellular donor concentration (M)	Required energy (ATP-equivalent)
Phosphorylation	S, T, Y / (D, H)	ATP / GTP	4.67E-03 (ATP) / 6.77E-04 (GTP)	1
Acetylation	K, (N' amine group)	Acetyl-CoA	2.88E-05	2
Methylation	R, K / (C, D, Q)	Methionine	6.39E-04	2
N-glycosylation	N	Dolichol-linked glycan	(Data deficit)	2 (per each sugar molecule)
O-glycosylation	S, T	UDP-GluNAc, UDP-GluNAc, etc.	8.97E-03 (UDP-GluNAc)	2
Hydroxylation	K, P, S	2-oxoglytarate	7.97E-04	4
S-nitrosylation	C	Nitric oxide	1.00E-09	6~7
Ubiquitination	K	E2-linked ubiquitin	8.50E-05	2 (per each ubiquitin molecule)
Sumoylation	K	E2-linked SUMO	(Data deficit)	2 (per each SUMO molecule)
Long-chain acylation	C, S, M, (N'-terminal G)	Palmitoyl-CoA, Myristoyl-CoA, etc.	1.0~8.5E-05	2

Table 1. Common PTM types found in eukaryotes

PTM types	Number of modified sites in human (SWISS-PROT)	Number of modified sites in human (dbPTM)	Number of non-redundant sites (PhosphoSitePlus)	Number of modified human proteins (SWISS-PROT)
Phosphorylation	40665	44704	171284	8214
Acetylation	6610	14222	38152	3406
Methylation	2360	5097	18614	989
N-glycosylation	15983	1691	6398	4408
O-glycosylation	1110	1092	3828	428
Hydroxylation	1264	67	(Data deficit)	148
S-nitrosylation	66	750	(Data deficit)	57
Ubiquitination	635	34159	70755	2435
Sumoylation	5894	1453	7568	1514
Acylation	708	170	(Data deficit)	847

Table 2. Number of known PTM sites and affected proteins in human proteome

H), cysteine (Cys, C) and so on. While different amino acids such as histidine are often favored in prokaryotic systems (22), the majority of phosphorylation events occur on S/T/Y residues in eukaryotes (23). More specifically, phosphoserine and phosphothreonine are much more abundant than phosphorylate in cellular environment, with the ratio of pS:pT:pY = 1800:200:1 (24).

Protein phosphorylation is by far the most common PTM in eukaryotic proteome (18) (Table 2). Multiple databases indicate about 40% of PTM annotations are associated with protein phosphorylation, which are distributed across at least 8,000 different proteins in human (25) if not more than to-third (11). In addition, more than 500 protein kinases and 150 protein phosphatases are directly involved in the phosphorylation of side chain (9), which not only allows more substrates to be recognized by kinase, but also enables cells to tightly regulate the phosphorylation status of proteins.

There are several properties of phosphorylation which might explain why the protein phosphorylation is widespread. First, chemical group donor of protein phosphorylation is ATP (or rarely GTP), which is among the most common metabolite species inside the cell (26). In chemical kinetics, reaction rate of all additive PTMs follow this equation

$$\mathrm{r} = k_{PTM}\,[\mathrm{Substrate}][Enzyme][Donor]^{n} \quad \ldots (1)$$

Here, r is reaction rate, k_{PTM} is reaction rate constant for given PTM, [Substrate] / [Enzyme] / [Donor] are concentrations of individual components, and n is number of donors required for given PTM, which is 1 for most of PTM types (27). Therefore, high concentration of ATP ensures substrates to be phosphorylated quickly: compared to other donor molecules, only UDP-acetylglucosamine, a chemical group donor for O-glycosylation, has higher cellular concentration than ATP (Table 1) (26).

Second, protein phosphorylation requires minimal amount of energy. Phosphorylation of single S/T/Y requires one ATP molecule, which doubles as an energy source. On the other hand, while many of PTM types do not require additional energy source during the reaction itself, regeneration of donor molecules require at least two ATP molecules, meaning more energy is needed for these processes (Table 1).

Third, protein phosphorylation modifies interactive capacity of side chain. Double negative charge and

multiple free electron pairs of phosphate groups allow extensive salt bridge / hydrogen bond network with other amino acids or solvents, thereby drastically affecting the local environment. Salt bridge formed between phosphorylated amino acid and positively charged lysine / arginine is notable as it is one of the strongest non-covalent interactions formed between amino acid (28), which is strong enough to induce conformational shift and its consequences (29). Also, phosphate group may adopt three charge states, -1/-2/-3, which allows it to affect local energy landscape in multiple ways.

Moreover, phosphorylation events are almost invariably reversible, which means modified protein could revert to its pre-modification state if necessary (13). All these properties make phosphorylation a suitable option for regulation of biological processes which requires immediate responses and reversibility, such as cell signaling, environmental responses or transcription factor regulation.

There are numerous examples of biological processes affected by phosphorylation. Receptor tyrosine kinases and MAP kinase pathway allow cell to recognize extracellular ligands and change transcription pattern accordingly (30). Protein kinase A (PKA), which the activity is dependent on cellular cAMP level, could provoke large-scale changes in cellular metabolism (31). CAM kinases (CAMK) are activated by the increase of calcium ion concentration in the cell and produce numerous effects, including cytoskeletal reorganization and neural cell growth (32). Protein kinase C (PKC), which is activated by not only calcium ion but also small metabolites such as diacylglycerol (DAG), is responsible for smooth muscle cell contraction (33). Different isoforms of casein kinase 1 (CK1) are involved in generating circadian rhythm, and mutation of these are associated with sleep disorders (34).

Despite its significance, only a small fraction of phosphorylation sites have been individually studied. Among hundreds of thousands of possible phosphorylation sites identified with mass spectrometry, only about 12,000 sites have been individually validated (25, 35), and even smaller fraction of those are annotated with identified functions or corresponding kinases. This leaves properties and biological implications of the majority of phosphorylation sites in obscurity.

[1-3] Proline-directed serine/threonine (S/T-P) phosphorylation sites

Multiple phosphorylation sites often share a common sequence element, or sequence motif (36). Active sites of PTM-inducing enzymes have unique three-dimensional arrangements of side chains, which make them to bind to substrates with specific patterns around the modified side chain. Sequence motif is well known example of patterns which enzymes recognize, and based on this, phosphorylation sites within proteins only known by its sequences could be predicted (37). The earliest phosphorylation site predictors were based on sequence motifs of phosphorylation sites (38).

The most commonly found sequence motif around eukaryotic phosphorylation sites is proline residue at +1 site of phosphorylated serine/threonine. These S/T-P sites are often referred to as 'proline-directed' sites, as accompanying proline is crucial in enzyme-substrate binding. In human, about 33% of phosphorylated serines and 47% of phosphorylated threonines fall in S/T-P category, which makes up about 1/3 of all phosphorylation sites identified so far. On the other hand, proline is avoided around phosphorylated tyrosines, which is also remarkable (Figure 1).

There are many examples of S/T-P sites which are crucial for regulation of biological processes. The most prominent example is transcriptional regulation: downstream kinases of MAP kinase pathway (e.g. ERK1/2, JNKs) phosphorylate S/T-P sites in transcriptional activators (e.g. c-FOS, Elk1) or supressors (e.g. Erf1) and change transcription pattern accordingly (39). Nuclear receptors such as glucocorticoid receptor (GR) have multiple S/T-P sites which are targeted by different kinases and produce different effects when phosphorylated (40). Nuclear kinases (e.g. NIMA) and phosphatases (Cdc25) often have S/T-P sites which modulate enzyme activity, which consequently affect transcription patterns indirectly via changing phosphorylation status of other nuclear proteins (41). In fact, S/T-P sites tend to appear more frequently in nuclear proteins, especially in transcription factors (42). On the other hand, S/T-P sites in cytoplasmic proteins are involved in different aspects of cell physiology. Phosphorylation of thr286 in cyclin D1 is associated with ubiquitination and subsequent degradation, which leads to G1 phase arrest (43). S/T-P phosphorylation sites in tau protein are implied to be involved in hyperphosphorylation-induced aggregation behaviors (17).

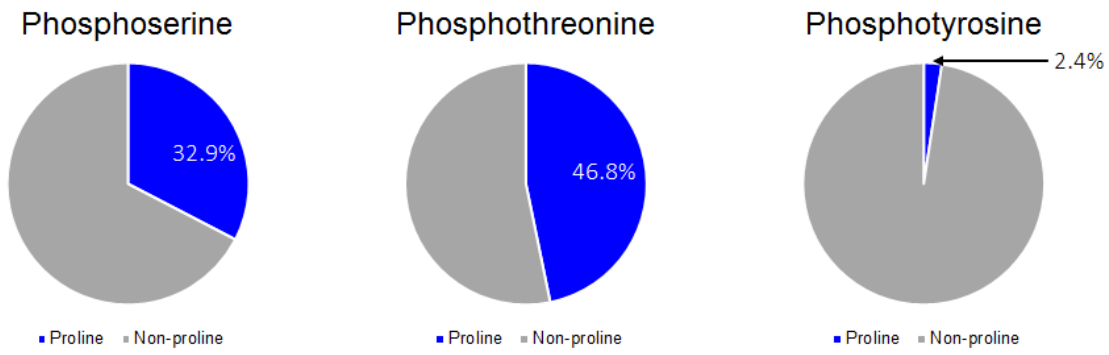

Figure 1. Frequency of phosphorylation sites with +1 proline residue

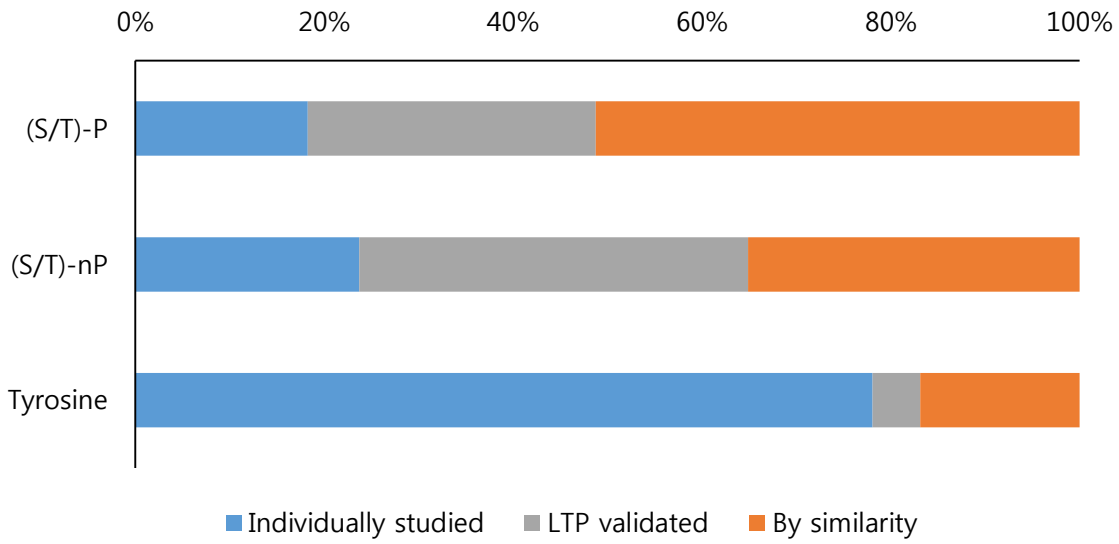

Figure 2. Phosphorylation sites by annotation status

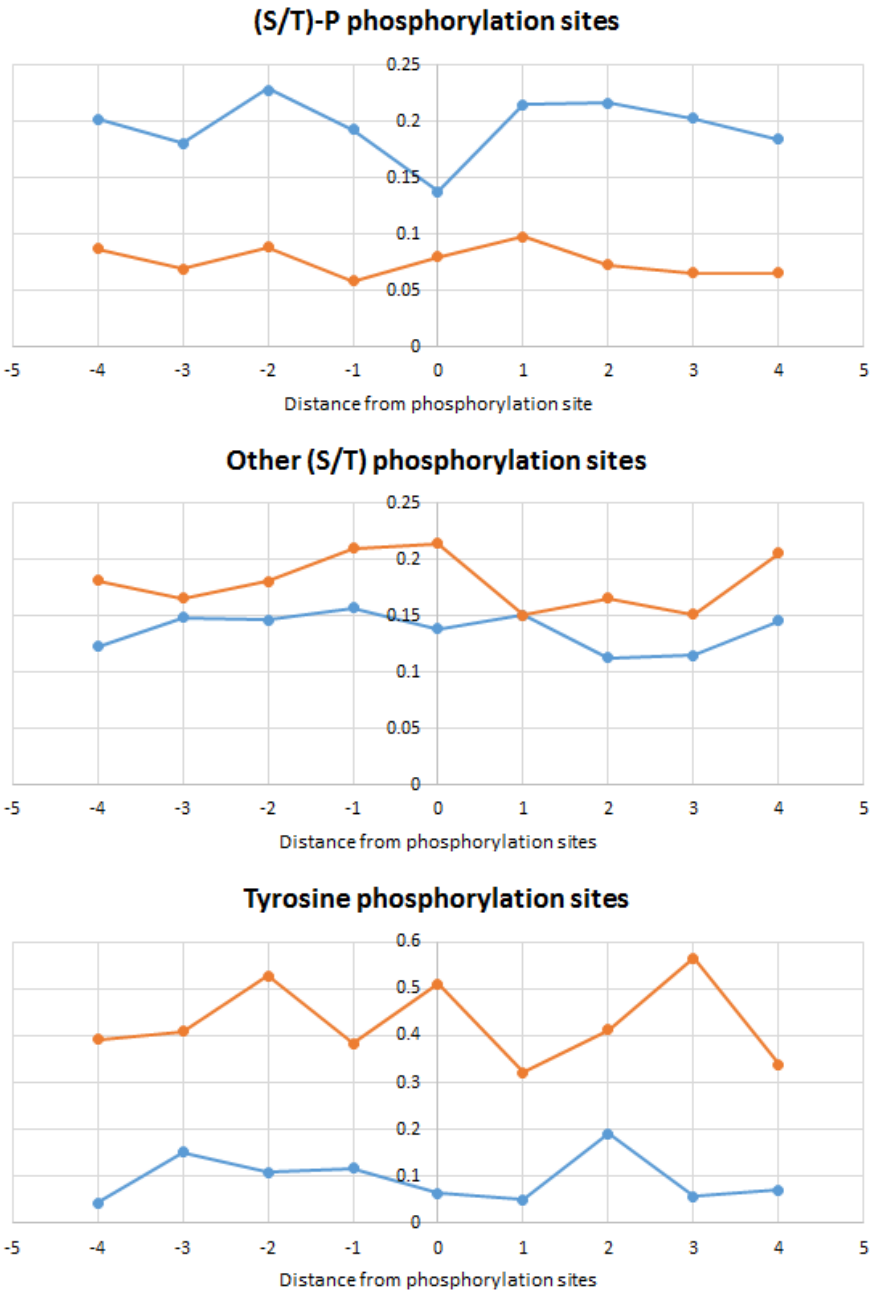

Figure 3. Frequency of clinical variants around phosphorylation sites being benign (blue) and pathogenic (orange)

However, when compared to other phosphorylation sites, information regarding S/T-P sites is sparse. According to SWISS-PROT and PhosphoSitePlus (25, 35), probabilities of being individually studied were 18.2% for S/T-P sites, 24.1% for other serine / threonine phosphorylation sites, and almost 80% for tyrosine phosphorylation sites. This was also consistent with the percentages of phosphorylation sites annotated by similarity to similar sequences in SWISS-PROT, which is 51.2% / 35.1% / 16.9% respectively for three categories (Figure 2).

Moreover, according to the clinical data from ClinVar database (44), it was found that mutations around S/T-P sites were less likely to produce pathogenic phenotypes. Missense mutations within ±4AA distance from the S/T-P phosphorylation sites have less than 10% probability of being pathogenic, including the phosphorylation site itself (thereby removing status as a phosphorylation sites). Compared to ~20% and ~50% probabilities found for other serine / threonine phosphorylation sites and tyrosine phosphorylation sites respectively, probability of missense mutations being pathogenic is significantly lower around S/T-P subclasses. At the same time, probability of missense mutations being benign was the highest around S/T-P phosphorylation sites (Figure 3).

This brings an uncertainty whether the individual S/T-P site has a discernible, independent and site-specific functionality or it is just an artifact of promiscuous kinases which is just tolerated within phosphoproteome. Nevertheless, S/T-P sites as a whole are indeed pivotal for normal physiology. Knockout mouse phenotype data shows targeted knockout of CMGC kinases likely produce lethal (either embryonic or neonatal) phenotypes as other kinases, showing that these are at least as essential as other protein kinases (45). This brings another possibility that the individual phosphorylation sites may have negligible functions by itself, but either as a part of phosphorylation site cluster or as a whole phosphoproteome targeted by a kinase, they may cause significant changes in the cellular environment.

[1-4] Biophysical properties of proline

For the fundamental understanding of properties of S/T-P sites, understanding of proline must precede: *Sensu*

stricto, proline is unique as it is the only imino acid among proteinogenic amino acids. Alkyl side chain of proline is connected to backbone nitrogen and forms a pyrrolydine ring structure, which consequently removes amide hydrogen from resulting polypeptide.

Side chain of proline itself is non-polar, which makes proline to be classified as aliphatic or hydrophobic, amino acid in classical hydrophobicity scales. However, absence of amide hydrogen poses local imbalance in backbone hydrogen bond donor / acceptor relationship, which makes sequestering proline residue from the surface to be energetically unfavorable (46). Compared to valine, which has the same number of side chain carbons and similar molecular weight, proline is far less likely found in the hydrophobic core of globular proteins and often found on protein surfaces (47). In addition, solubility of polyproline is higher than other peptides such as polyglycine, polyalanine and polyleucine, largely due to its lack of amide hydrogen (48). For these reasons, despite its non-polar side chain, proline effectively behaves as a hydrophilic amino acid.

Proline is generally referred to as a disruptor of secondary structures for good reasons. Because of its lack of amide hydrogen, proline residues cannot form a hydrogen bond which is often crucial for stabilizing conformations such as alpha helix or beta sheets (36). Formation of ring structure, on the other hand, imposes a steric constraint in phi dihedral angle of proline: proline could only have phi angles between -110° ~ -30° (49). For this reason, proline is almost absent in beta sheets, which has an average phi value of -140° for antiparallel sheets (which is more common) and -120° for parallel sheets (which is less common) (49). In contrast, proline is often enriched at the N'-terminal of alpha helices, as it could function as a cap which stabilizes alpha helix energetically (50). Also, proline-rich peptides may adopt a polyproline II helix conformation, a special type of left-handed helix structure which have no in-between hydrogen bonds and highly exposed backbone.

Another important property of proline is that it could adopt cis- conformation frequently. Most of peptide bond is trans- peptide bond: the probability of cis- conformation is less than 0.1%. However, in Xaa-Pro peptide bond, free energy difference between cis- and trans- isomers is relatively smaller: due to steric conflict between alpha carbon of preceding amino acid and delta carbon of proline, free energy of trans isomer is

relatively high, which allows around 5% of Xaa-Pro bonds to be in cis- state (51). For this reason, along with glycine, proline is often a necessary component in formation of turns (e.g. beta turns) which requires cis-conformation. At the same time, the activation energy of cis/trans isomerization is up to ~20 kcal/mol (52), which means spontaneous conversion between cis- and trans- states is very slow (in biochemical sense). This makes proper isomerization of proline to be a pivotal step in protein folding.

Proline could be referred to as either rigid or flexible amino acid, depending on the perspective. As its ring structure not only limits possible phi angle but also interferes with preceding amino acid, proline itself has an inherent conformational rigidity. On the other hand, most of prolines are actually found in intrinsically disordered regions which do not adopt any stable three-dimensional structures, suggesting proline inside a polypeptide is strongly associated with local flexibility. This dual nature allows proline to occupy specialized (but also pivotal) roles in biological systems; proline-rich domains often allow IDPs to have significantly higher hydrodynamic radii than expected (53), while it could also promote peptide compaction via formation of cis- isomers (54); coupled with charged amino acids, proline residues prevent stacking of beta sheets and thereby curb amyloid-like aggregation behaviors (55); high proline and glycine contents allow elastomeric proteins, such as elastin, to be extended in response to mechanical stress and recoiled without being denatured (56).

Beside of phosphorylation, proline is also associated with many PTMs. Modification of proline itself is not really diverse: only (2S/4R)-4-hydroxyproline, or simply hydroxyproline (Hyp) occurs frequently in human. The fact should be noted is, Hyp is more abundant than seven proteinogenic amino acids (Cys, Gln, His, Met, Phe, Trp and Tyr), which makes non-reversible proline hydroxylation to be one of the most common PTM in human (57). Hyp is associated with several roles; free energy gap between cis- and trans- isomers of Xaa-Hyp is relatively large, which means cis- isomer is less favored in Hyp (58); Hyp in collagen (and putatively collagen-like proteins) stabilizes triple helices (59); hydroxylation of HIP-1a induces ubiquitination of HIP-1a by von Hippel-Lindau ligase complex and subsequent degradation (60). On the other hand, PTMs such as proteolysis and N-/O-glycosylation are found to be associated with adjacent proline residue. For example, N-glycosylation is enhanced with -2 proline and hindered with either -1 / +1 proline. O-glycosylation is not

associated with proline at specific location, but generally favors high frequency of prolines nearby, particularly at -1 and +3 sites (61). Interestingly, O-glycosylation, which also modifies serine and threonine, often competes with phosphorylation and known to produce either similar (62) or opposite (63, 64) conformational effects.

[1-5] CMGC kinases

S/T-P sites are known to be strongly associated with a specific family of kinases called CMGC (65). This kinase family is named after its main subgroups: cyclin-dependent kinases (CDKs), mitogen-activated protein kinases (MAPKs), glycogen synthase kinases (GSKs) and CDK-like kinases (CDKLs). With about 60 members, CMGC kinase family is one of the major kinase groups in eukaryotes (66).

As the name implies, CMGC kinases are crucial components in multiple signaling and regulation pathways. CDKs are classically associated with cell cycle progression (e.g. Cdk1, Cdk2, Cdk4) but many of those are involved in transcriptional regulation which is not directly associated with cell cycle (e.g. Cdk9, Cdk12) (67). MAPKs, as noted above, are responsible for phosphorylation of numerous transcription factors and cause large-scale changes in transcription patterns. GSK3 is known to target 'primed' – already phosphorylated – substrates and promotes not only glycogenesis but also modulates hundreds of important downstream substrates. Other notable members such as dual-specificity kinases (DYRKs) (68), serine/arginine-rich protein-specific kinases (SPRKs) (69) and homeodomain-interacting protein kinases (HIPKs) (70) play pivotal roles in transcriptional regulation and mRNA splicing.

CMGC kinases share similar active site architectures which allow enzymes to recognize substrates with proline. In CDK2, Val164 has an unusual left-handed conformation which gives binding site a pocket structure devoid of electron donors. As hydrogen bond formation is impossible in this pocket, only proline, which does not have amide hydrogen, can fit into this pocket and allow neighboring residue to be phosphorylated (71). ERK2, a MAP kinase, has a similar structure formed by Val186 and Ala187 (72). However, in GSK3, pTyr216 (which is cognate with pTyr185 in ERK2) is moved away from the substrate

which makes pocket structure to be more open: this makes GSK3 to be significantly less specific towards S/T-P sites (73). There is another research conducted in yeast system which the mutation of active site changes preferred +1 residue from proline to arginine (74).

The problem is, the number of CMGC kinases compared to S/T-P sites is relatively low, which makes each kinase to recognize much more substrates than any other kinases. Assuming each phosphorylation is targeted by single kinase (which is not true), each CMGC kinase should recognize and catalyze ~180 protein substrates: this number is significantly higher than ~60 for other serine/threonine kinases and ~20 for tyrosine kinases (Figure 4). At the same time, sequence motifs of CMGC kinases are often poorly defined: in extreme cases (e.g. Cdc2, Erk1, p38), no amino acids other than phosphorylated serine/threonine and accompanying proline is enriched around phosphorylation site, which strongly implies the promiscuity of kinases (75). This might be consistent with the fact that kinases such as Cdk1, GSK3B, JNK1, and MAPK1 affect hundreds (if not thousands) of substrates (76).

[1-6] PIN1 prolyl isomerase

Proline is the only amino acid which could frequently adopt cis conformation (36). However, due to high activation energy of isomerization, speed of spontaneous cis-trans isomerization is very slow (37). (Peptidyl) prolyl isomerases bind to peptides with proline and lowers activation energy by reducing partial double bond nature of peptide bonds, thereby facilitates transition from one isomer to another (38).

Prolyl isomerases, which are often referred to as catalytic structural chaperones (78), are evolutionarily conserved family of enzymes which could significantly alter protein functionality by inducing large-scale conformational changes which often result in modified interaction patterns. PIN1 isomerase is unique among prolyl isomerases as it specifically targets peptide bond between phosphorylated S/T-P sites (79). PIN1 has N'-terminal WW domain and C'-terminal catalytic domain: WW domain, which is named after two conserved tryptophan residues in sequence, have high binding affinity towards phosphorylated S/T-P. Isomerization of S/T-P sites is important as different binding partners, including kinases and phosphatases, only interact with

cis or trans isomer, but not both (80). As phosphorylated S/T-P has slower rate of isomerization that non-phosphorylated S/T-P, facilitation by isomerase is often crucial for regulation of biological pathway in timely manner.

PIN1 is known to catalyze hundreds of different substrates in human, and each isomerization would bring different molecular consequences. In some substrates, isomerization enhances protein stability and prevent degradation (e.g. p53, beta-catenin, c-Fos), while the opposite happens in different substrates (e.g. c-Myc, SRC-3, GRK2). Isomerization by PIN1 is also known to induce translocalization of substrates (e.g. Cyclin D1, NF-kB) or regulate PTM status of other sites in same protein (e.g. Tau, Raf-1) (81).

This allows PIN1 to be involved in numerous biological pathways, which means misregulation of PIN1 would result in pathological phenotypes. For example, PIN1 is highly expressed in multiple types of cancer cells, and is correlated with poor prognosis (79). PIN1 affects numerous transcription factors and other mitotic proteins which allow cells to develop cancerous phenotypes such as resisting cell death, genomic instability, modified energy metabolism and evasion from immune responses (79). Also, In Parkinson's disease, PIN1 indirectly promotes Lewy body formation by binding to synphilin-1 and affects alpha-synuclein to aggregate (82). On the other hand, in Alzheimer's disease, significant decrease of PIN1 expression level is observed, and knockout of PIN1 is sufficient to reproduce Tau/Amyloid beta related pathological phenotypes (83). PIN1 interacts with Tau, which has multiple S/T-P sites, and promotes dephosphorylation (84) and possibly degradation (80), thereby reduce the chance of forming aggregation. For this reason, multiple upstream regulators of PIN1 exist and these tightly control the expression level and catalytic activity of PIN1 (79).

There are still many questions about mechanism of PIN1. While it is classically assumed that WW domain recruits substrate to the catalytic domain, recent researches have suggested that while WW domain prefer trans isomer, catalytic domain usually brings cis to trans isomerization, which is conflicting with previous understandings (85). There are many alternative explanations to this discrepancy, but none is backed with substantial evidences. Also, while hundreds of substrates and its sequence information are known, it is unclear whether PIN1 have an extra ‘preference’ towards certain sequence element or not. As there is no

evidence of S/T-P sites which are not interacting with PIN1, it is possible that S/T-P phosphorylation and consequent isomerization is a universal mechanism, which affects all conserved S/T-P motif and exerts function by inducing conformational shift.

All aforementioned evidence suggests that S/T-P phosphorylation is both widespread and biologically important. It is also associated with distinct subset of enzymatic machineries and possible biophysical mechanism which could not take place in other types of phosphorylation sites, making it an intriguing subject of research. However, despite it is distinct characteristics, S/T-P sites did not get emphasized as frequently as other phosphorylation sites, especially when compared to tyrosine phosphorylation sites, and its niche in biological system is still not clearly understood.

[1-7] Intrinsically disordered protein (IDP) / Intrinsically disordered regions (IDR)

After the first protein structure was elucidated by X-ray crystallography, a 'structure-function' paradigm - specific three-dimensional structure determines biological function of molecule - has been dominating the field of biology (86). Coupled with Anfinsen's dogma, which states amino acid sequence of protein determines its three-dimensional structure, the paradigm evolved into sequence-structure-function paradigm and became a framework of structural biology (87). In this perspective, protein sequence encodes all of the information about structure and function, meaning features and properties of given protein could be deducted from it (88).

While it is still largely true for many of proteins, over the past two decades, this paradigm has been challenged by the concept of intrinsically disordered region (IDR) of protein. IDR is a region of protein which has no fixed three-dimensional structure in the physiological condition: instead, IDR exist in a conformational ensemble, a continuum of multiple structural states allowed by its free energy (89). If there is no fixed structure entirely, the protein is referred to as intrinsically disordered protein (IDP).

IDR is a common element in the proteome. About 35% of human proteome sequence is predicted to be intrinsically disordered. More than half of proteins are predicted to have at least one IDR, while the proteins

falling in IDP category is quite rare (90). Also, the abundance of IDR show a roughly positive correlation between proteome size and biological complexity, meaning amount of IDR has increased much faster than the expansion of proteome itself (89).

Having no fixed structure does not mean IDR/IDP have no biological function: on the contrary, IDR is a crucial component of the eukaryotic proteome. Dunker and colleagues distinguished 28 functions for disordered regions, which could be categorized into six groups (91, 92). The most basic example would be entropic chains, including flexible linkers connecting structured domains into a single polypeptide, which function due to its flexibility. Another simple one is display sites, such as linear motif recognized by other molecules: these sites requires to be exposed to surface and devoid of unfavorable interactions. The assemblers, which bring multiple binding partners and induce high-order complex formation, form another important group of IDRs.

On the other hand, ensemble nature of IDR allows allosteric regulation of protein, which transfers thermodynamic change at one site to another and consequently affects the function (93). The effectors, which not only binds to other proteins (as display sites) but also affects the activity of binding partner, would be an example of IDRs which the function is associated with allosteric regulation. Another two groups are molecular chaperones, which bind to multitude of proteins and facilitate its proper folding, and scavengers, which sequester small ligands and neutralize it.

More specific examples of IDR include hub proteins which could bind multiple interaction partners at once and allow those to interact with each other (94); histone tails which harbors PTM sites and control chromatin status and/or gene transcription level (95); ER chaperones such as calnexin and calreticulin which the C'-terminal IDR is crucial for substrate binding and subsequent facilitation of folding (96); N'-terminal domains of steroid hormone receptors and nuclear receptors which could both recruit interaction partners and allosterically regulate ligand binding region (LBD) and/or DNA binding region (DBD) (97); p21 and p27, which are folded upon binding to interaction partners such as cdk2/cyclin A and inhibit its kinase activity (98); nuclear localization signal or other trafficking signals which induces intracellular translocation (99); and so on.

[1-8] IDRs and protein phosphorylation

For the last two decades, researchers have found that various PTM types are associated with IDRs (100). For example, phosphorylation sites, methylation sites, glycosylation sites and ubiquitination sites are overrepresented in IDR sequences, while SUMOylation sites, myristoylation sites or modified cysteines are more likely to be found on folded regions (101).

Phosphorylation is one of the first PTMs which the association with intrinsically disordered region is recognized (102) and also the most thoroughly studied. At least 50% of phosphorylation sites are on the regions predicted to be IDR, while the frequency is just around 30% for non-phosphorylated serines and threonines, or even lower for non-phosphorylated tyrosines (Figure 5). Preference of phosphorylation sites towards IDR could be interpreted in two ways: first, phosphorylation sites are enriched in IDRs as IDRs are mostly exposed to the solvent (display site function). In physiological environment, protein phosphorylation is not known to happen spontaneously: it always requires other kinases which catalyze transfer of phosphate group from chemical donors. This requires phosphorylation sites to be presented to the outside of protein and become accessible to kinases. IDRs are frequently associated with low hydrophobicity, high charge and high flexibility, which all promotes exposure to solvent and consequently provides a suitable environment for kinase binding (103). This also allows phosphoserine / phosphothreonine / phosphotyrosine amino acids to be exposed and binds with other interaction partners which specifically recognize phosphorylated substrates.

On the other hand, phosphorylation sites might be enriched in IDRs as the chemical modification of IDRs would produce larger effects than the same modification of folded region (effector function). Ensemble nature of IDR conformation allows small chemical change introduced by PTM to significantly affect conformational equilibrium, which is almost impossible in folded structure where the free energy difference between native state and other conformational states is large (104). For example, hydrodynamic properties of IDRs are sensitive to changes in electrostatic environment: meaning PTMs which change side chain charge, such as phosphorylation, could substantially affect the dynamics of protein (105). Also, structural propensity values of phosphoserine, phosphothreonine and phosphotyrosine are different from its non-phosphorylated

counterparts, which might read to either stabilization or disruption of local conformations (106). These two effects could also trigger large-scale changes involving not only IDRs but also folded regions, thereby functioning as a 'switch' of protein behaviors.

The conformational effect may manifest in many different forms; phosphorylated residues could function as a N'-terminal cap of alpha helix, which reduces free energy of alpha helical state and subsequently promotes its formation (62); when coupled with negative charges and extension-promoting amino acids, phosphorylation induce PII conformation, a common conformation adopted by peptide ligands (107); phosphorylation of serine residues in FUS protein by DNA-PK interferes with phase separation behavior of FUS and disassembles droplets (108); Ser66/76 double phosphorylation of p19INK4d induces local unfolding and dissociation of inhibitory complex, ultimately releases CDK6 and initiates S-phase entry (109); conversely, Thr37/46 double phosphorylation of 4E-BP2 induces disorder-to-order transition which reduces affinity to eIF4E and enables interaction with eIF4G (110). These varying consequences suggest the conformational effects of phosphorylation are highly dependent on its surrounding environment and associated interaction partners.

Besides, there are other examples of roles of phosphorylation sites in IDR regions which are not directly associated with conformational ensemble. For instance, phosphorylation of Ser139 of variant histone γ-H2AX appears after DNA damage and recruits DNA repair proteins and signaling factors, halts transcription / translation and induce chromatin relaxation (111); phosphoserine / phosphothreonine allows nearby serine / threonine to be recognized as a valid ('primed') substrate of GSK3B, which results in a variety of consequences (112); phosphorylation of DNA-binding domains typically reduces its binding affinity towards DNA (113); localization signal sequences often include serine / threonine phosphorylation sites which either enhance or impede translocation when phosphorylated (114)

[1-9] Aim of this work

The ultimate aim of this project was to understand the biological significance of proline-directed serine/threonine (S/T-P) phosphorylation sites. To fulfill this aim, we approached in two ways: first, we analyzed amino acid sequences and biophysical properties of phosphorylation sites and demonstrated S/T-P phosphorylation sites form a distinct subclass which is statistically separated from other S/T phosphorylation sites. We further validated our findings by incorporating these findings into a new phosphorylation site prediction algorithm, PHOSforUS, which relies on simple framework but outperforms currently available phosphorylation site predictors. Second, we analyzed ortholog sequences of known human and mouse phosphoproteins and found evidences supporting both enrichment of S/T-P phosphorylation sites and different patterns of evolution in mammals.

Following sub-objectives are associated with pre-stated approaches.

- To show that the criteria which separate S/T-P phosphorylation sites from non-phosphorylated SP/TP dipeptide are different from those for other S/T phosphorylation sites: Chapter 2

- To show different biophysical properties of S/T-P phosphorylation sites would cause different consequences after phosphorylation: Chapter 3

- To show identified biophysical properties could be utilized to construct a phosphorylation site predictor with superior predictive performances: Chapter 4

- To show +1 proline is evolutionarily conserved - more likely to predate phosphorylated S/T residues, and the change between phosphorylation site subclasses is a rare event: Chapter 5

[2] S/T-P phosphorylation sites form a distinct subclass within serine/threonine phosphorylation sites

[2-1] Introduction

In this chapter, I'd like to demonstrate in detail that S/T-P phosphorylation sites form a distinct class within phosphorylation sites.

There are two main classes of eukaryotic phosphorylation sites people acknowledge - serine / threonine (S/T) phosphorylation sites and tyrosine (Y) phosphorylation sites. Tyrosine phosphorylation sites are not only different from the other class by its target residue, but also by its sequence preference, associated kinases, intracellular localization of substrate proteins and biological functions of substrate proteins (115). In addition, tyrosine kinase family emerged in eukaryotic level, while it is presumed that the common ancestor of eukaryotic kinases predates it (116), implying biological processes involving tyrosine kinases would have emerged later on.

My strategy here is to draw parallels between tyrosine phosphorylation sites and S/T-P sites; (1) S/T-P sites have different sequence characteristics from both other S/T phosphorylation sites and tyrosine phosphorylation sites; (2) S/T-P sites are predominantly targeted by specific family of kinases which seldom recognizes substrates without +1 proline; (3) Phosphoproteins with S/T-P sites show different expression and localization patterns from other phosphoproteins; (4) Phosphoproteins with S/T-P sites are associated with different molecular functions and biological processes. The evidence will thereby indicate that S/T-P phosphorylation sites are not a mere subpopulation of S/T phosphorylation sites with an easily recognizable marker, but a separated 'class' of phosphorylation sites with distinct properties and functionalities.

[2-2] Approaches

-2.2.1. Classification of phosphorylation sites

Based on current understandings of phosphorylation sites, I devised five-class / three-class classification

scheme of phosphorylation sites. Along with classical division based on modified residue itself, I added an extra criterion based on the presence of +1 proline to divide serine and threonine phosphorylation sites.

A five-class model classifies phosphorylation sites into S-nP, S-P, T-nP, T-P and tyrosine classes. On the other hand, three-class model merges serine and threonine phosphorylation sites into a single group, but still divides it by the presence of +1 proline, thereby classifies phosphorylation sites into S/T-nP, S/T-P and tyrosine classes. I used five-class model for the most of part of my research, but also utilized three-class model to assess generalizable characteristics of S/T-P phosphorylation sites.

-2.2.2. Data sources

Canonical human protein sequences were obtained from SWISS-PROT (25), a manually curated subset of the UniProt database. Phosphorylation annotations were obtained from SWISS-PROT and PhosphoSitePlus (35, 148).

Phosphorylation site datasets were assembled from SWISS-PROT annotations and low-throughput (LTP) subset of PhosphoSitePlus. Sequence fragments of 29 amino acids (14 residues N-terminal and C-terminal relative to a central phosphorylation site) were extracted from these sets and subsequently divided into five subsets (S-P, S-NP, T-P, T-NP, Y) based on the identity of the center residue and the presence of Pro as its C-terminal neighbor (148). Non-phosphorylated sequence datasets were generated by removing all possible phosphorylation sites from SWISS-PROT and PhosphoSitePlus (both LTP and HTP). Resulting statistics of these datasets are shown in Table 3.

Lists of phosphoproteins with specific phosphorylation class (‘inclusive’) and phosphoproteins ONLY with specific phosphorylation class (‘exclusive’) are generated from canonical protein sequences and assembled phosphorylation site datasets. Resulting statistics of six datasets are also shown in Table 3.

Class	Phosphorylation sites	Non-phosphorylated sites	Class	All phosphoproteins with specific phosphorylation class ('inclusive')	Phosphoproteins only with specific phosphorylation class ('exclusive')
S-P	10348	30170	**S/T-P**	4792	975
T-P	2688	20943			
S-nP	21936	455303	**S/T-nP**	6251	2208
T-nP	3045	288492			
Tyrosine	2058	145170	**Tyrosine**	1032	229

Table 3. Statistics of phosphorylation site datasets & phosphoprotein datasets

Human kinase-substrate pair correspondence information was collected from Phospho.ELM (117). List of kinases and kinase family information was retrieved from Uniprot (https://www.uniprot.org/docs/pkinfam). Protein abundance level information was collected from PaxDB (118), which includes protein abundance information from 169 different human cell lines and tissues.

-2.2.3. Sequence logo creation

Phosphorylation class-specific Sequence logos were created using Seq2Logo 2.0 tool (119) by using PSSM option. Position-specific weight matrices were created for both phosphorylation site datasets and non-phosphorylated datasets by dividing observed site-specific amino acid frequency by average frequency of human proteome. In the sequence logo, polar amino acids are colored green, negatively charged amino acids are colored red, positively charged amino acids are colored blue and proline is colored purple.

Amino acids which are observed more than expected appear above the baseline while those less than expected appear below the baseline. Sizes of each letter denote the significance of enrichment / depletion of that amino acid at given site.

-2-2-4. Kinase-substrate relationship analysis

From collected kinase-substrate pair information, I attempted to demonstrate the bilateral relationship of phosphorylation site classes and kinase families in two ways. First, I calculated how many of specific class of phosphorylation sites are catalyzed by a specific kinase family, to see the dependence of phosphorylation site class on kinases. To address the opposite side of question, I calculated class-specific frequency of substrates for each individual kinases and kinase families. I manually curated the notable kinases - kinases which shows atypical frequency of specific phosphorylation class - to discover whether there is a feature which make these kinases remarkable.

-2.2.5. Expression pattern analysis

Expression levels of proteins were calculated from 169 protein abundance datasets. Each dataset not necessarily contains values for every canonical proteins, which could mean either the protein is non-detected or dataset is incomplete. For the missing values, I assigned small base value - 1/100 of the minimum abundance value found from all datasets to avoid possible computational errors. Average expression levels of proteins were calculated by taking a geometric mean of abundance values. From these values, 8 datasets were generated - 2 for whole phosphoprotein and whole non-phosphorylated protein respectively and 6 for phosphoproteins with specific phosphorylation class (S/T-nP, S/T-P & tyrosine / 'inclusive' and 'exclusive').

-2-2-6. GO enrichment analysis

GO term enrichment analysis was done with tools from Gene Ontology Database (120, 121), which is in turn connected with PANTHER database (122). I applied 9 protein datasets - 6 for phosphoproteins and 3 for kinases (CMGC family, tyrosine kinase family, other S/T kinases) - and calculated GO term enrichment for three aspects – biological process, cellular localization (compartment) and molecular function.

Resulting fold enrichment values and p-values were analyzed with custom python script and visualized as a color map which outlines significant GO terms. Identified GO terms were manually curated to screen out duplicative terms and compared between classes to identify the differences.

[2-3] Results

-2,3.1. S/T-P sites have distinct sequence features

From the sequence logos for phosphorylated / non-phosphorylated sequences, I could characterize the specific features of each phosphorylation classes. First, serine / threonine phosphorylation sites (other than S/T-P sites) are clearly characterized by enriched charged amino acids near phosphorylation sites (Figures

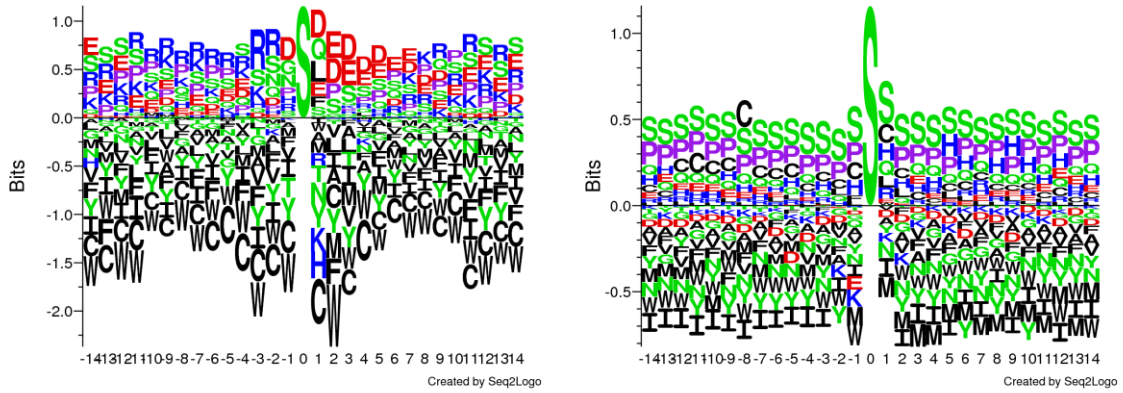

Figure 4. Average sequence landscape of S-nP phosphorylation sites (left panel) & corresponding non-phosphorylated sequences (right panel)

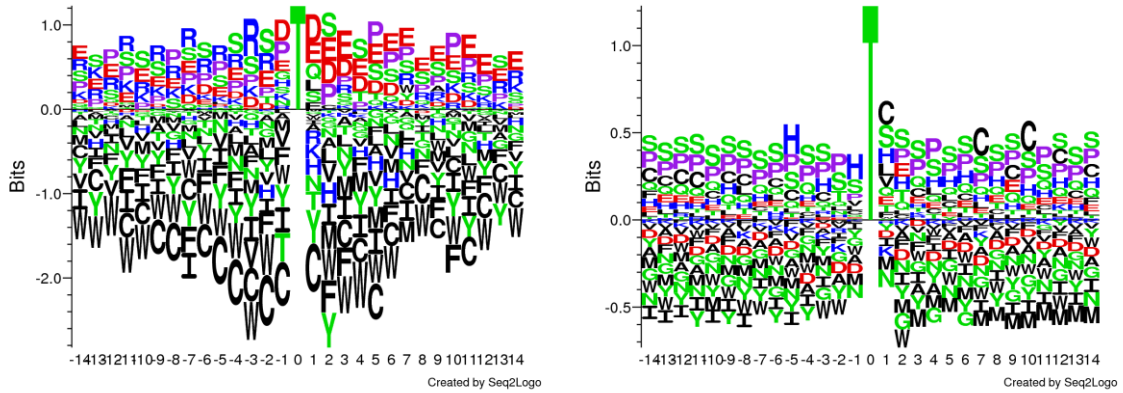

Figure 5. Average sequence landscape of T-nP phosphorylation sites (left panel) & corresponding non-phosphorylated sequences (right panel)

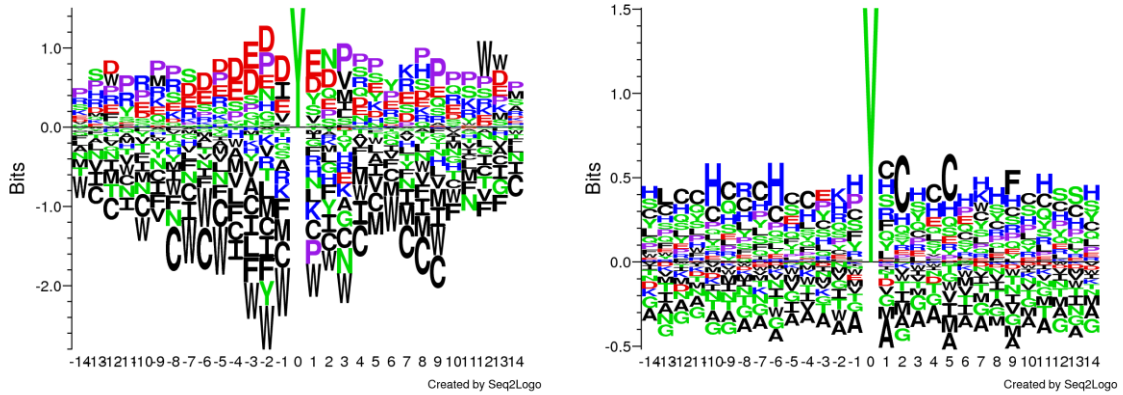

Figure 6. Average sequence landscape of tyrosine phosphorylation sites (left panel) & corresponding non-phosphorylated sequences (right panel)

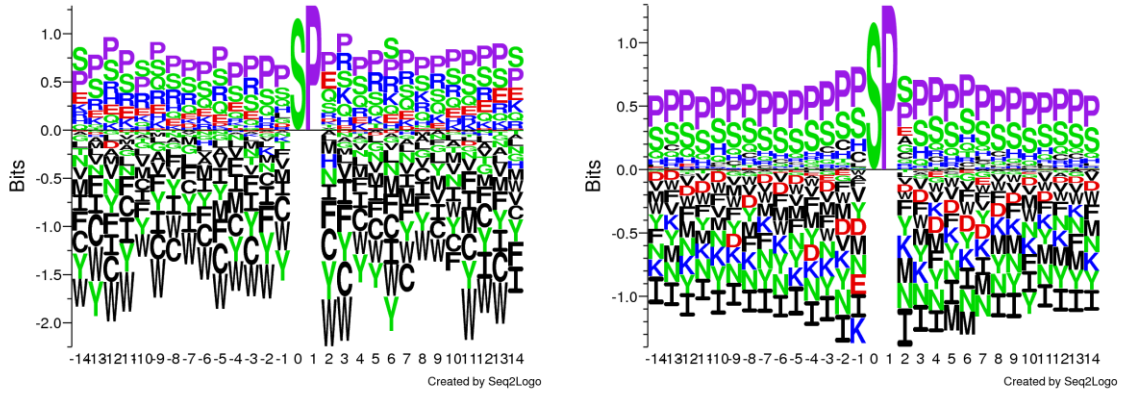

Figure 7. Average sequence landscape of S-P phosphorylation sites (left panel) & corresponding non-phosphorylated sequences (right panel)

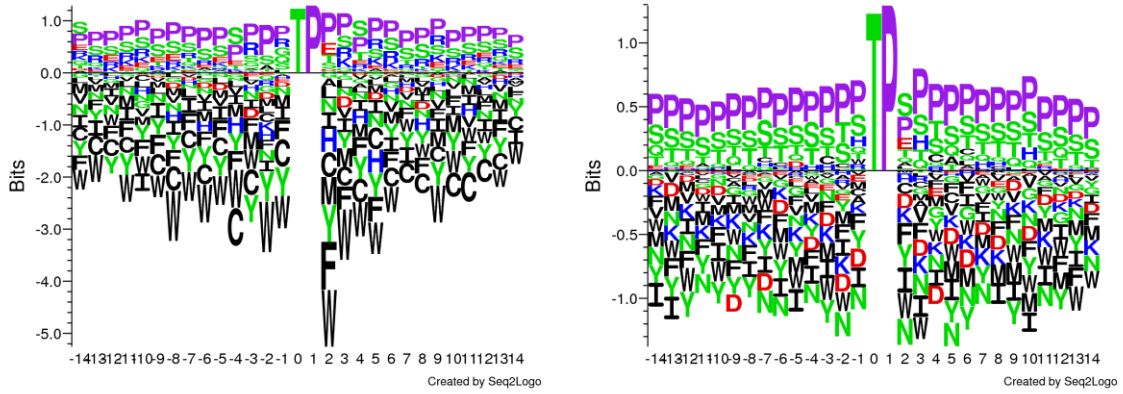

Figure 8. Average sequence landscape of T-P phosphorylation sites (left panel) & corresponding non-phosphorylated sequences (right panel)

Family name	Members	Representative	Target residue	Known motifs
AGC	58	PKA, PKG, S6 kinases	S/T	-3/-2 basic residues (PKA, PKG), proximal (-3~+3) basic residues (PKC), +1 hydrophobic residue
CAMK	75	CamK, MLCK, AMPK, EF2K	S/T	-3 Arginine, proximal (-3~+3) basic residues
CK1	12	CK1, VRK2	S/T	-3/-2 basic residues (universal)
CMGC	62	CDK, MAPK, GSK	S/T	+1 Proline (universal), -2 Proline (MAPK), -3/-2 basic residues (CDK), +4 phosphorylated residue (GSK3)
NEK	11	NEK1	S/T	(-1/+2 arginine) or (-1 phosphotyrosine & -5/-4/-2 acidic residues)
RGC	5	RETGC, NPR	S/T	-
STE	56	MAPKK, MAPKKK, STK	S/T/Y	N'-terminal basic residues (PAK), C'-terminal basie cesidues (MST)
TK	90	ABL1, EGFR, JAK	Y	High frequency of acidic residues (especially N'-terminal), -1/+1/+3 hydrophobic residues
TKL	34	RAF1, RIPK, SKR3	S/T	N'-terminal basic residues, +2/+4 serines
Others	80	AURK, CK2, PKR, PLK1		-

Table 4. List of eukaryotic kinase families found in human proteome

Family name	Members	Representative	Target residue	Known motif
ADCK	5	ADCK1	(Not clear)	(Possibly small molecule kinases)
Alpha	6	ALPK1, TRPM7	S/T	Alpha-helical conformation
FAST	1	FASTK	S/T	
PDK	5	PDK1, BCKDK	S/T	Proximal threonine residues (-3, -2, +4)
PI3/PI4	9	ATM, ATR, mTOR, PIK3	S/T	+1 glutamine (ATM/ATR), Proximal (-2~+2) hydrophobic residues (especially +1) (mTOR)
RIO	3	RIOK2	S/T	

Table 5. List of atypical kinase families found in human proteome

4, 5). It is notable that the distribution of charge is asymmetric – N'-terminal side (-4 ~ -1) is enriched with positively charged amino acids (K/R) while C'-terminal side (+1 ~ +5) is enriched with negatively charged amino acids (D/E). On the other hand, frequencies of cysteine, histidine & proline residues are significantly decreased around modified residue. This is consistent with the previously identified substrate consensus sequences of kinases (Table 4, 5). For instance, kinases in AGC or CAMK kinase family recognize peptides with K/R at -3 site as a proper substrate, while casein kinase II or ATM/ATR kinases strongly prefer substrates with D/E at +1~+3 sites. However, it should be noted that sequence logos were generated using every phosphorylation site sequences fall in certain category, meaning each of the individual phosphorylation sites is not likely to have both types of residues around the modified serine / threonine.

Tyrosine phosphorylation sites show strong preference towards negatively charged amino acids around (-7 ~ +2) the modified tyrosine (Figure 6). Mild increase of frequencies of aliphatic residues (I/V) and decrease of frequencies of positively charged amino acids (K/R) and cysteine (C) are also observed nearby phosphorylation sites. Similarly, this is consistent with previously identified consensus sequences of tyrosine kinases. Interestingly, most of aforementioned 'positive sequence markers' are charged amino acids, which suggests proper distribution of charges is crucial in kinase-substrate binding process. This is supported by the fact that those consensus charged amino acids are more likely to be found in N'-terminal side of phosphorylation site. Due to the domain architecture of eukaryotic protein kinases, substrate binding domains of kinases usually interact with N'-terminal side of phosphorylation sites (123) and recognize its composition – which justifies accumulation of positive markers on N'-terminal side.

On the other hand, sequence logos generated from S/T-P phosphorylation sites were very similar to those generated from non-phosphorylated counterparts (Figures 7, 8). While subtle increase of positively charged amino acids and decrease of high-molecular weight amino acids overall, it is not sufficient to be mentioned as a positive marker of substrates required for kinase recognition. Also, it is notable that the proline residue is the most frequently found amino acids nearby (-5 ~ +5, not only +1 site) phosphorylation sites, which is not true for other phosphorylation site classes. Again, this is consistent with previously identified consensus sequences of CMGC kinases: beside of universal +1 proline and some subfamily-specific markers such as -

3 K/R for CDKs or -2 proline for MAPKs, consensus motifs of individual CMGC kinases are generally poorly defined.

The data shows S/T-P phosphorylation sites are placed in different sequence environment from those for other types of phosphorylation sites. This implies the potential differences in involved biophysical mechanism, associated interaction partners and physiological functions, which I will elaborate one-by-one in the later sections.

-2.3.2. S/T-P sites are modified by a specific family of kinases

It is already known that S/T-P sites are associated with CMGC kinases (73) but there is no detailed analysis of the relationship between these two. How many of S/T-P sites are actually phosphorylated by one of CMGC kinases? How many of substrates targeted by specific CMGC kinases fall in S/T-P category? How the other types of phosphorylation sites interact with S/T-P sites? To address these questions, I analyzed currently known kinase-substrate correspondence information, which provides both substrate sequence information and targeting kinase identity: by classifying substrates by residue types, it was able to better analyze the relationship between specific kinase family and substrate class.

Among 3,233 kinase-substrate pairs collected from Phospho.ELM (117), 703 (21.7%) pairs were S/T-P sites (Table 6). Frequency of S/T-P pairs found was roughly consistent with frequency of S/T-P sites found in manually annotated phosphorylation sites (20.5% in SWISS-PROT, 22.4% in PhosphoSitePlus), which allowed me to assume that this dataset is not biased towards specific phosphorylation site class.

Among 703 S/T-P site pairs, 613 (87.2%) were pairs involving one of CMGC kinases (Figure 9). To be more specific, more than 90% of those were associated with either MAPK or CDK: only 39 pairs were associated with other CMGC kinases. On the other hand, none of other kinase families were associated with more than 5% of S/T-P site pairs: 22 (3.1%) for AGC kinases, and even lower for other kinase families. Considering only 35 (17.1%) out of 205 analyzed kinases are CMGC kinases, and only 667 (20.7%) out of 3223 pairs

Kinase family	Number of kinases	+1 Proline	Other residues	Total
AGC	23	22	822	844
Atypical:alpha	1	0	5	5
Atypical:PDK	5	0	10	10
Atypical:PI3/PI4	5	12	127	139
CAMK	24	10	228	238
CK1	4	2	45	47
CMGC	35	613	54	667
NEK	3	4	9	13
Others	17	21	381	402
STE	21	8	106	114
TK	52	9	703	712
TKL	15	2	40	42

Table 6. Statistics of Phospho.ELM kinase-substrate pair dataset

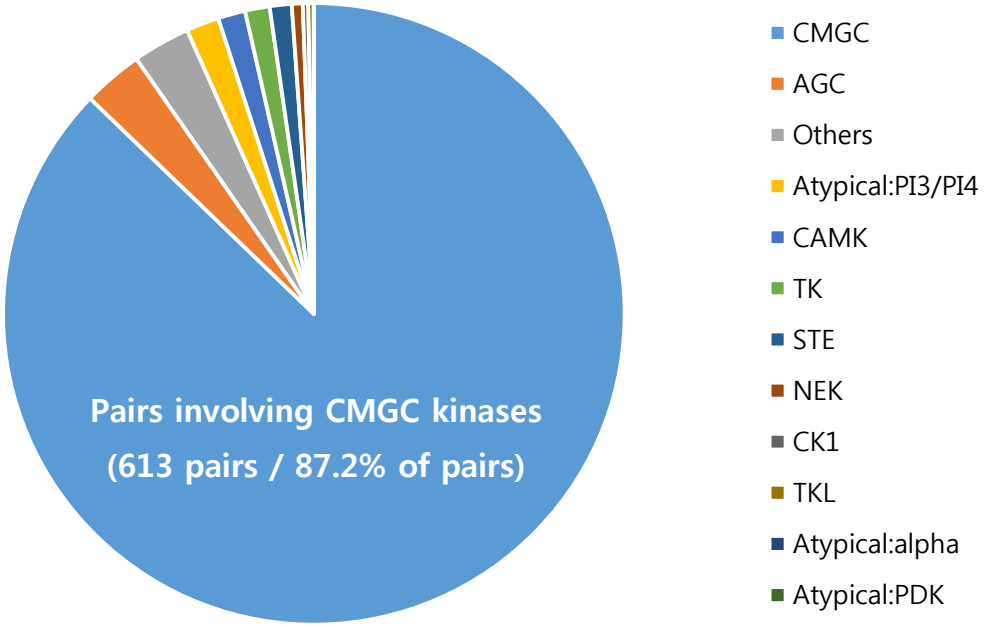

Figure 9. Kinase partners of S/T-P phosphorylation sites

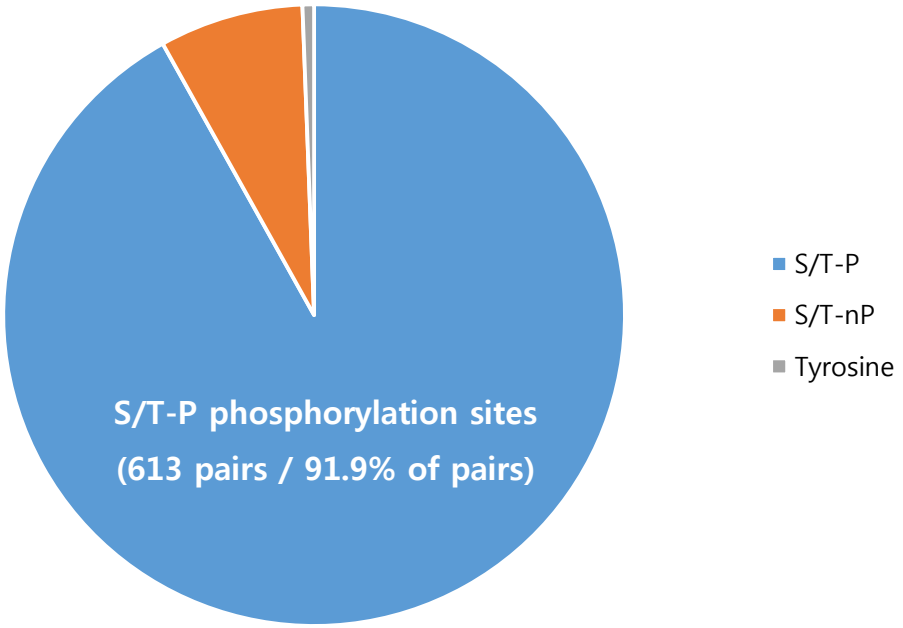

Figure 10. CMGC kinase substrates by phosphorylation class

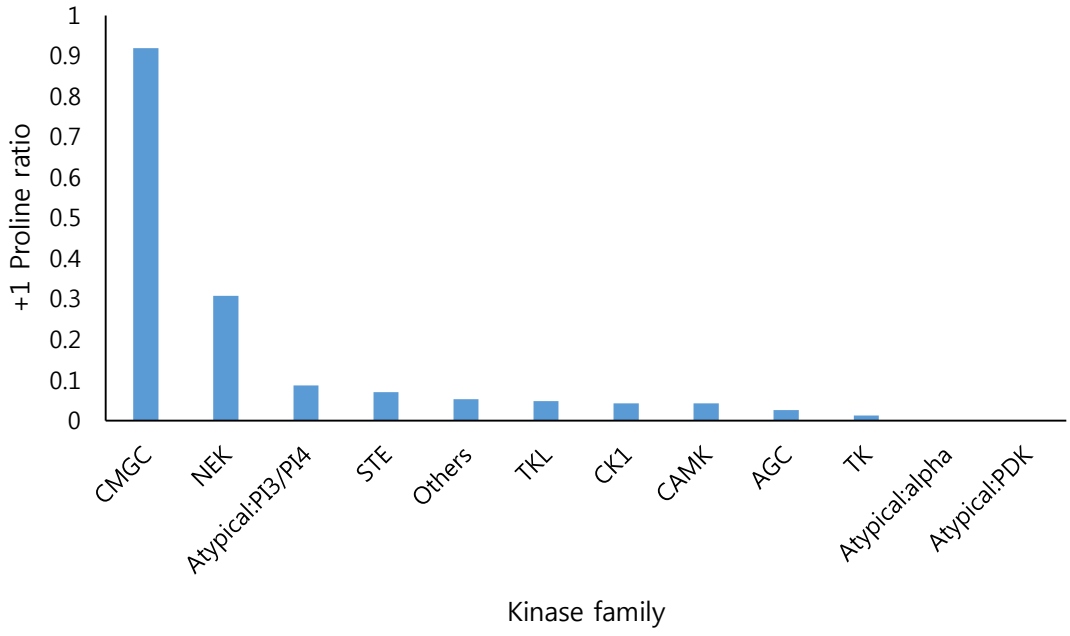

Figure 11. Ratio of substrates with +1 proline for each kinase family

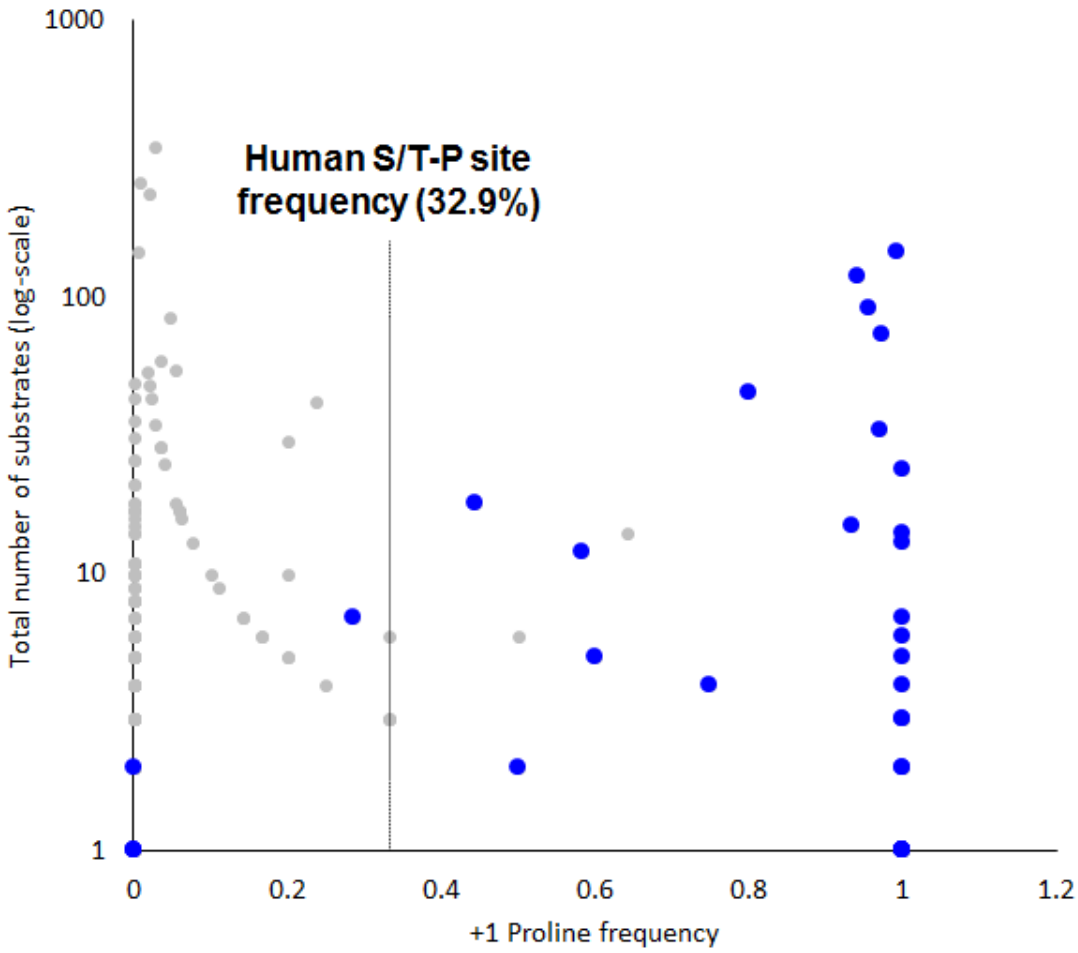

Figure 12. Kinase-specific +1 proline frequency (Blue: CMGC kinases, gray: non-CMGC kinases)

were associated with CMGC kinases, aforementioned 87.2% probability of being associated with CMGC kinases is indeed significant.

In kinase-side perspective, 613 (91.9%) out of 667 pairs associated with CMGC kinase family were S/T-P sites (Figure 10). This value is significantly higher compared to the frequencies of S/T-P substrates for major kinase families - 22 (2.6%) out of 822 pairs for AGC family, 10 (4.2%) out of 228 pairs for CAMK family, 9 (1.3%) out of 703 for TK family, and so on. NEK family, a small subgroup associated with 13 kinase-substrate pairs, showed high frequency (4 out of 13 = 30.8%) of S/T-P sites, but its statistical significance is not enough to be incontestable, especially when compared to CMGC family (Figure 11).

Frequency of S/T-P sites for each individual kinases shows a bimodal distribution (Figure 12) with modes at low S/T-P frequency (0~15%) and very high S/T-P frequency (85~100%). Here, 35 out of 39 kinases with frequency of S/T-P substrates higher than total S/T-P content (32.9%) were CMGC kinases, while 5 out of 166 kinases with frequency of S/T-P substrates lower than total S/T-P content were CMGC kinases. It should be noted that the exceptions found here actually had extraordinary characteristics, such as mTOR (S/T-P ratio = 64.3%), an atypical kinase which is not associated with eukaryotic kinase group, DYRK1B / DYRK3 (S/T-P ratio = 22.2% as a whole), dual-specificity CMGC kinases, and MAP2K7 (S/T-P ratio = 50.0%), another dual-specificity kinases but in STE family. This suggests that CMGC kinases are specialized enzymatic machinery to phosphorylate S/T-P sites, which are generally not preferred by other types of eukaryotic protein kinases.

Again, similar patterns are found with tyrosine phosphorylation sites and tyrosine kinases. None out of 712 pairs targeted by tyrosine kinases were either serine or threonine, and 712 out of 734 (97.0%) kinase-substrate pairs were associated with one or more of tyrosine kinases. Other 22 pairs were associated with either MAP2K subgroup of STE family (14 pairs), DYRK subgroup of CMGC family (3 pairs), ICK of CMGC family (1 pair), TGFBR2 of TKL family (3 pair) or WEE1 of non-classified eukaryotic kinases (1 pair). These kinases are previously noted for dual-specificity: an ability to phosphorylate both S/T and tyrosine (124).

The data clearly show that the majority of S/T-P sites are phosphorylated by a single kinase family, which is specialized to recognize +1 proline in the substrates, just as tyrosine phosphorylation sites are targeted by a

single kinase family. Another parallel could be drawn from here as CMGC kinases also emerged in eukaryotic level, suggesting the possibility that S/T-P phosphorylation is the another kind of intracellular regulation mechanism, just as tyrosine phosphorylation does.

-2.3.3. Proteins with S/T-P sites are expressed in a lower level

Along with the evidence I provided in previous sections, I'd like to show that each phosphorylation site classes are associated with characteristics of phosphoproteins, including expression pattern and intracellular localization. For this and later sections, I applied two different sample-selection schemes; first, I selected every phosphoprotein with a certain type of phosphorylation site, meaning there is an overlap between two sample groups ('inclusive' scheme); second, I selected phosphoproteins with only a specific type of phosphorylation sites, meaning there is no overlap but more than half of phosphoproteins are not included in any of three sample groups ('exclusive' scheme). Using both schemes would provide a better idea about the general functionality of each phosphorylation classes in the cell. The statistics of each sample groups are shown in table 3.

Figure 13 and 14 show that there is a significant difference of expression levels of phosphoproteins between sample groups. It was found that the average expression level of phosphoproteins was 2.75-fold higher than that of non-phosphorylated proteins, which implies biases towards highly expressed protein species posed by the limitations in current experimental techniques used for PTM research. Interestingly, the average expression levels of sample groups defined by the 'inclusive' scheme was 5.6-fold lower than those of groups defined by 'exclusive' scheme overall.

In 'inclusive' scheme, phosphoproteins with S/T-P sites show 1.6-fold lower (p-value = 7.35E-6) expression in average than those with other S/T phosphorylation sites, or 2.43-fold lower (p-value = 1.21E-12) than those with tyrosine phosphorylation sites. On the other hand, with 'exclusive' scheme, phosphoproteins

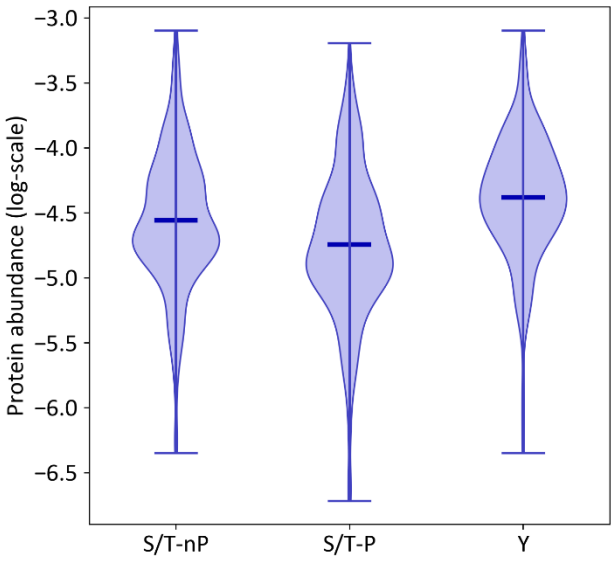

Figure 13. Average protein abundances of all phosphoproteins ('inclusive') with given phosphorylation sites

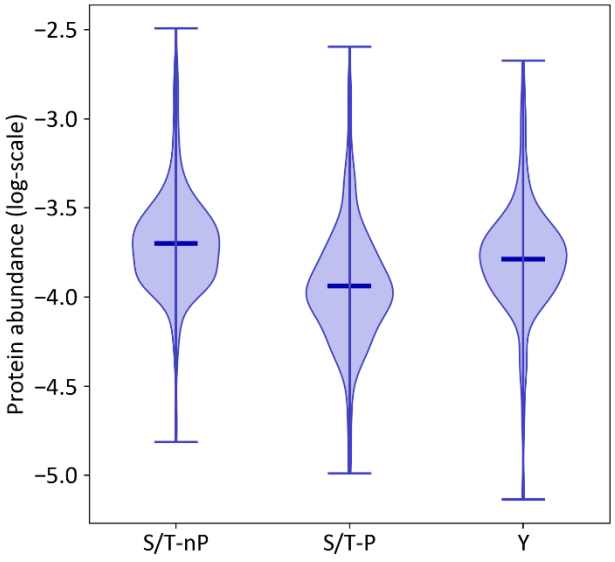

Figure 14. Average protein abundances of phosphoproteins which only contains ('exclusive') specific type of phosphorylation sites

with S/T-P sites show 2.2-fold lower (p-value = 1.55E-22) expression level than those with other phosphorylation sites, or 1.74-fold lower (p-value = 1.24E-6) than those with tyrosine phosphorylation sites. In any case, phosphoproteins with S/T-P sites were expressed in lower level, which is consistent with the lower concentration of protein species with IDRs (125) which many of S/T-P sites are associated. Interesting thing is, while the average expression level of all phosphoproteins with tyrosine phosphorylation sites was the highest among the sample groups defined by the 'inclusive' scheme, average expression level of phosphoproteins which only contains tyrosine phosphorylation sites was actually lower than that of phosphoproteins only with other S/T phosphorylation sites.

-2.3.4. Proteins with S/T-P sites show different intracellular localization patterns

Phosphoproteins with S/T-P phosphorylation sites were associated with a specific compartment within the cell – nucleus (Figure 15). GO enrichment analysis with the 'inclusive' scheme revealed that the highest-scoring compartments were nucleus (GO:0005634), nuclear lumen (GO:0005654), nucleoplasm (GO:0031981), organelle (GO:0043226), non-membrane bound organelle (GO:0043228), and so on (Figure xx). The enrichment pattern was quite similar between S/T-P group and S/T-nP group, while the difference was more obvious between S/T-P group and tyrosine group.

The difference was more obvious with the 'exclusive' scheme (Figure 16), which removes phosphoproteins which are affected by more than two classes of phosphorylation sites. Here, similarity between S/T-P group and S/T-nP group was diminished – suggesting that the similar-looking patterns found with 'inclusive' scheme were likely caused by the large number phosphoproteins which have both classes of phosphorylation sites (Figure 15). Along with the GO components I mentioned, intrinsic / integral component of membrane (GO:0031224 / 0016021), extracellular region (GO:0005576) and chromosome (GO:0005694 / 0000790 / 0000228) were also noteworthy. In contrast, enrichment scores were relatively lower for cytoplasm / cytosol (GO:0005737 / 0005829), plasma membrane (GO:0005886), cell periphery (GO:0071994), vesicle (GO:0031982), extracellular organelle (GO:0043230) and receptor complex

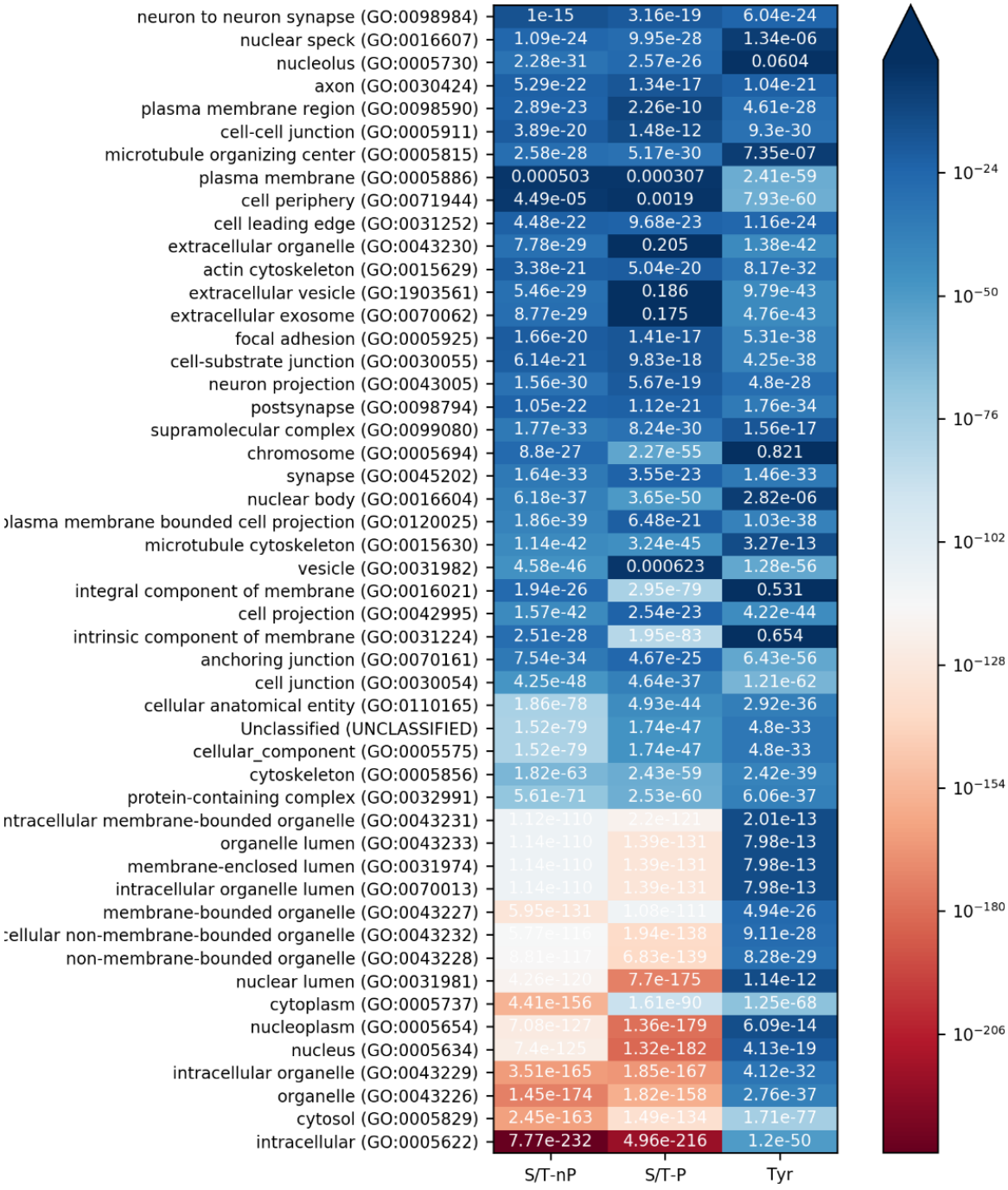

Figure 15. Cellular compartments enriched in specific group of phosphoproteins ('inclusive' scheme)

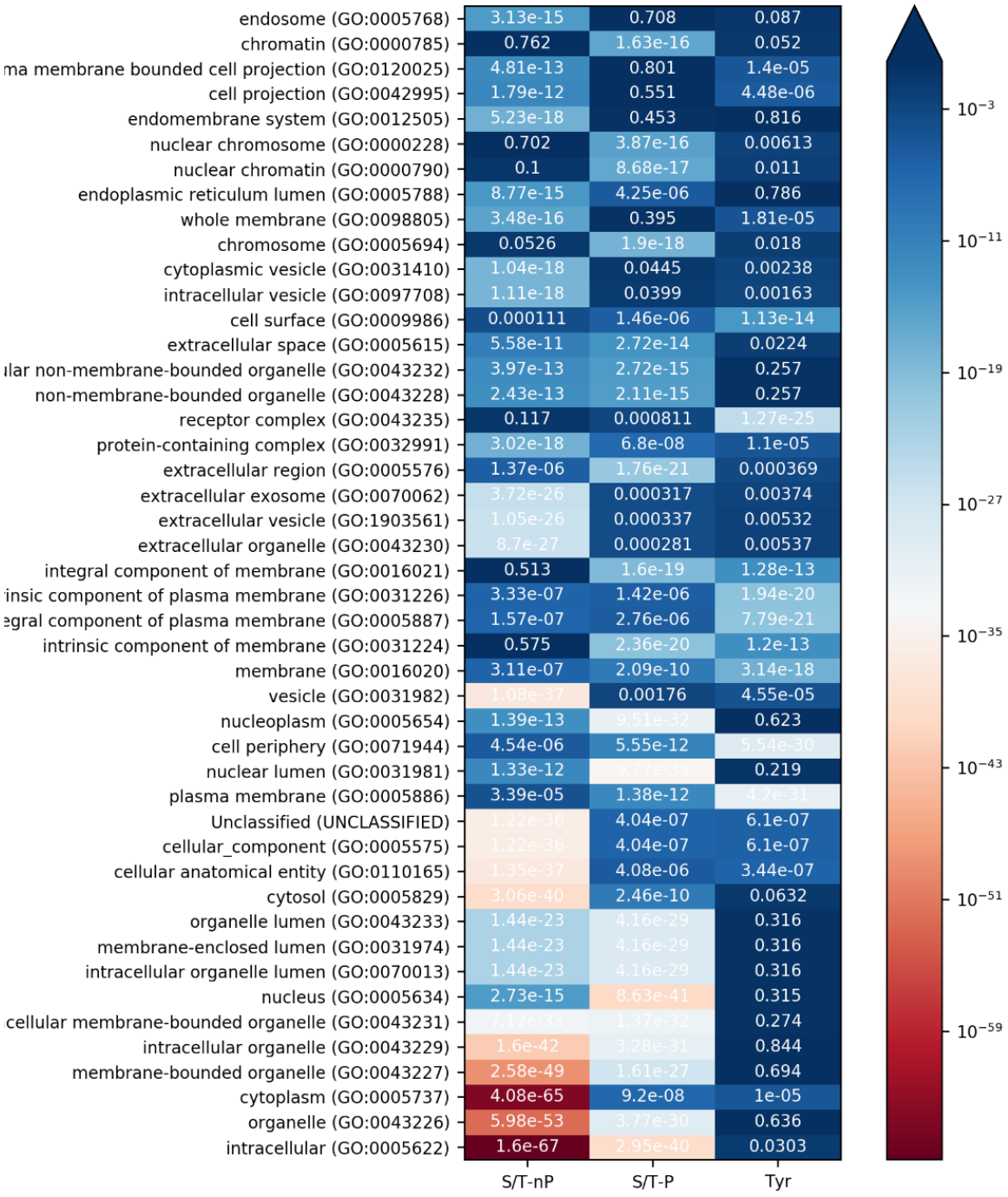

Figure 16. Cellular compartments enriched in specific group of phosphoproteins ('exclusive' scheme)

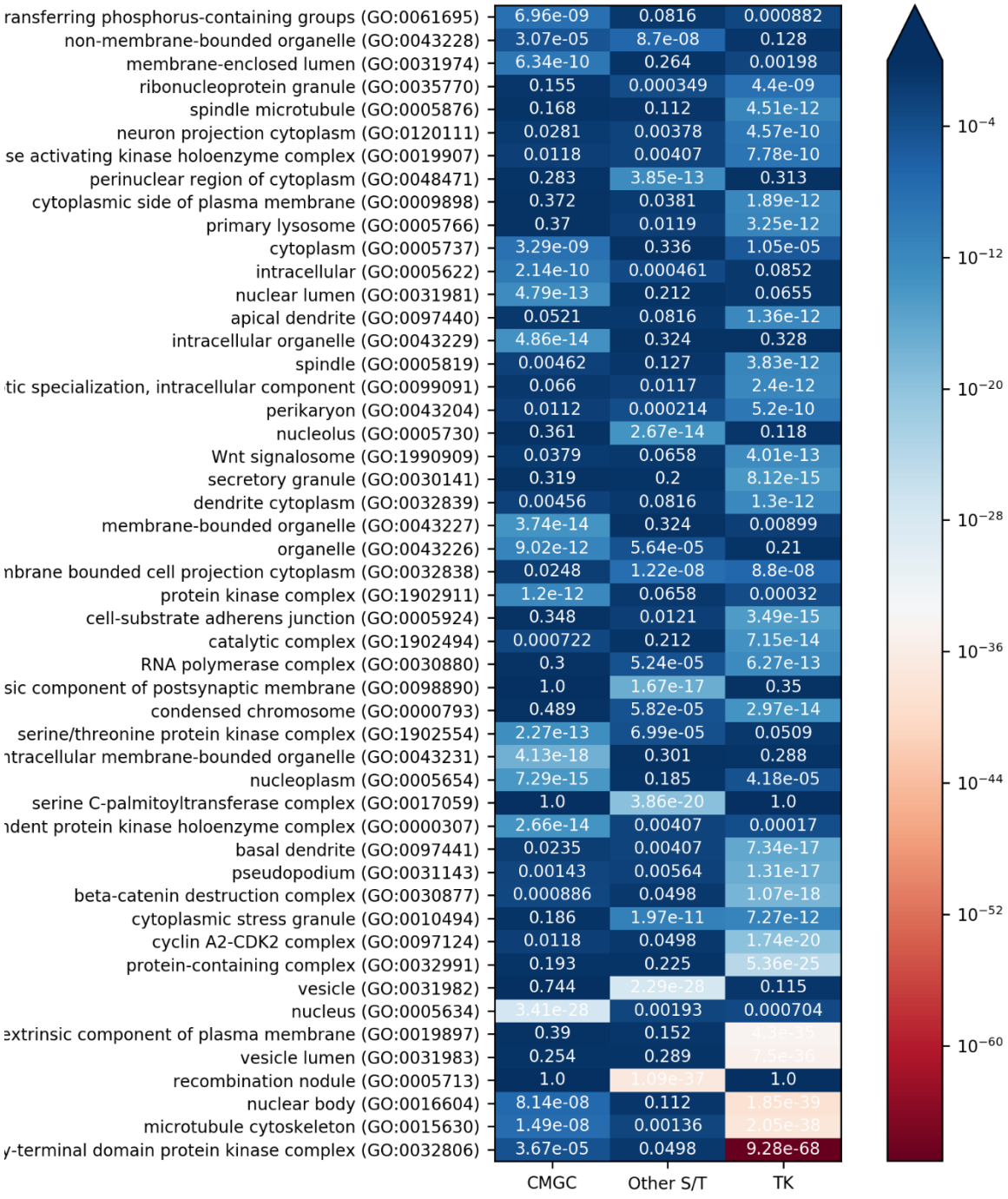

Figure 17. Cellular compartments enriched in specific group of kinases

(GO:0043225); these compartments were found to be strongly associated with either S/T-nP group or tyrosine group. On the other hand, phosphoproteins with S/T-nP sites were mainly associated with cytoplasmic components, including cytosol, vesicle, membrane-bound organelle (GO:0043227), cellular anatomical entity (GO:0110165) and extracellular organelle, while those were avoided disfavored in chromatin-associated compartments and receptor complexes. Phosphoproteins with tyrosine phosphorylation sites were strongly associated with membrane-associated compartments, including plasma membrane, cell periphery, membrane, receptor complex, intrinsic / integral component of plasma membrane (GO:0005887 / 0031226) and cell junction (GO:0031224).

The enrichment patterns found in phosphoproteins are consistent with those found in protein kinases (Figure 17). Ignoring protein complexes, CMGC kinases were localized in nucleus, nucleoplasm and membrane-bound organelles, while other S/T kinases show high association with vesicles, nucleolus (GO:0005730) and perinuclear region of cytoplasm (GO:0048471). GO terms associated with tyrosine kinases were highly associated with plasma membranes and membrane-associated receptors. This co-localization pattern would also support that the enrichment of proteins with certain phosphorylation site in specific cellular compartment is not coincidental, while it raises a question whether this enrichment pattern is a mere byproduct of kinase localization or genuine result of co-evolution.

-2.3.5. Proteins with S/T-P sites are associated with different biological functions

As a continuation from the previous section, I also found proteins with different types of phosphorylation sites tend to have different molecular functions (Figures 18~21). Molecular functions strongly associated with proteins with S/T-P sites include transcription regulator activity (GO:0140110), nucleic acid binding (GO:0003676), site-specific DNA-binding associated functions (GO:0000977) and cis-regulatory region binding (GO:0000978). On the other hand, biological process term strongly associated with proteins with S/T-P sites were mostly found to be nucleobase-containing compound metabolic process (which includes transcriptional regulation) (GO:0019219) and its subcategories.

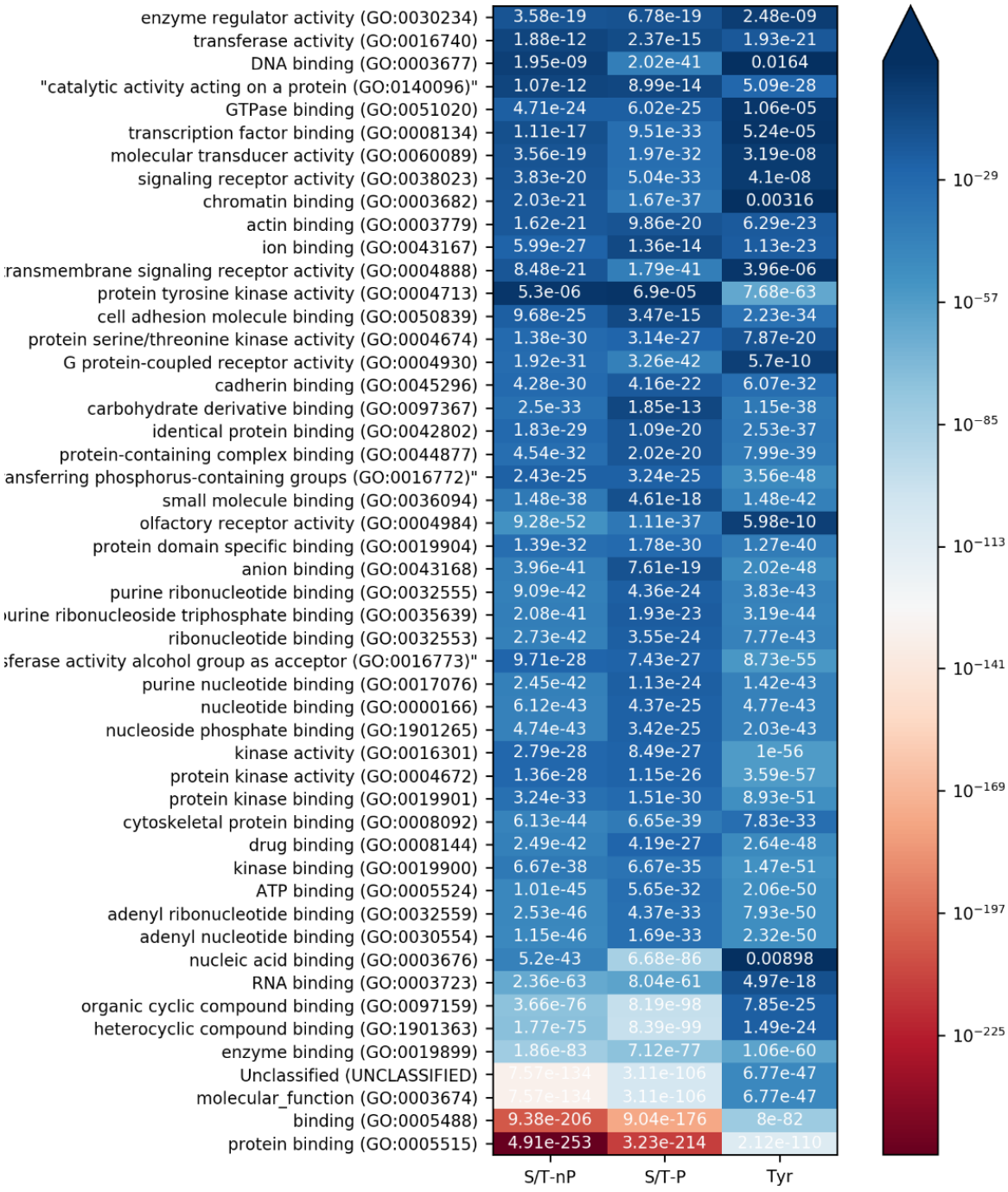

Figure 18. Molecular functions associated with specific group of phosphoproteins ('inclusive' scheme)

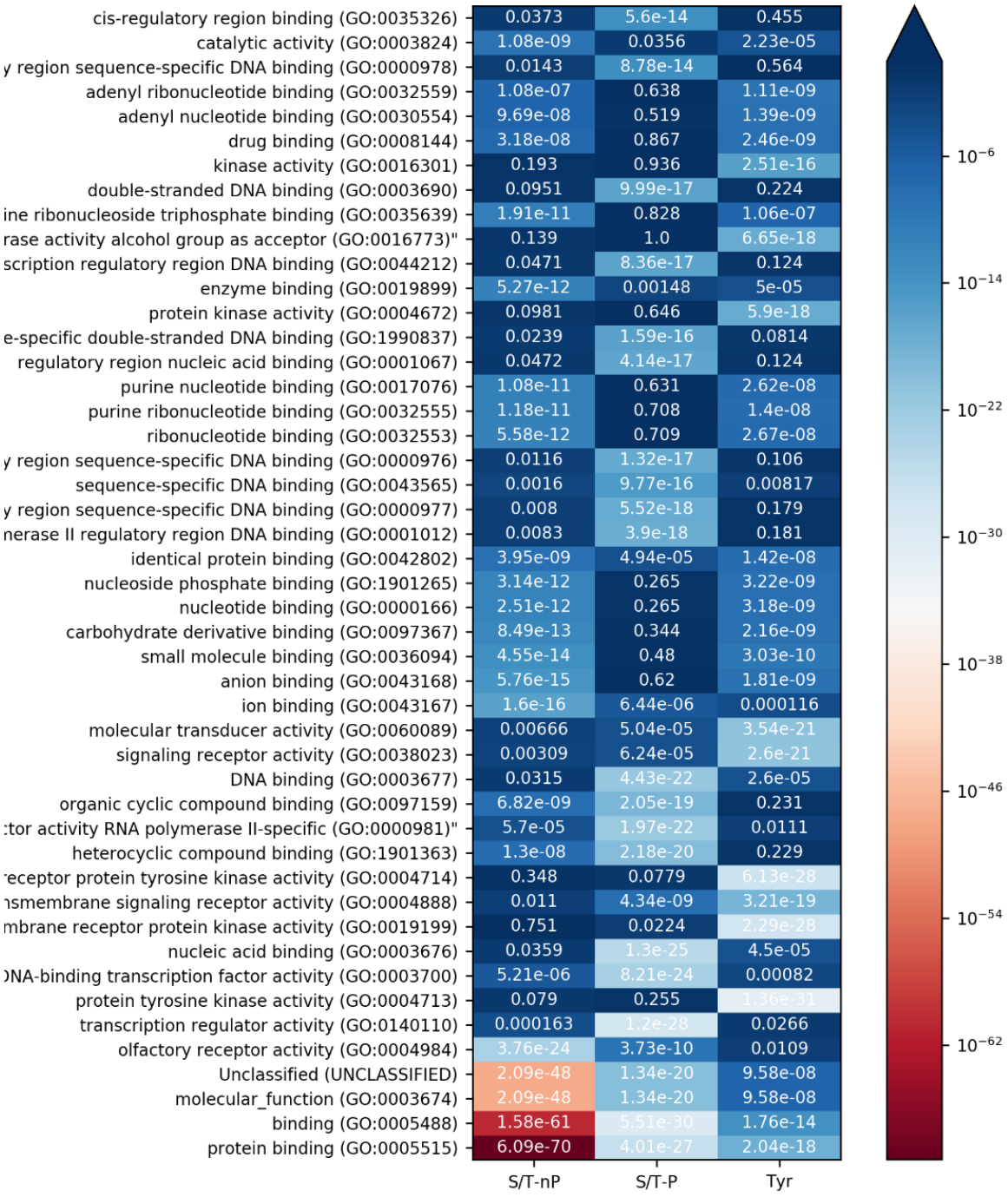

Figure 19. Molecular functions associated with specific group of phosphoproteins ('exclusive' scheme)

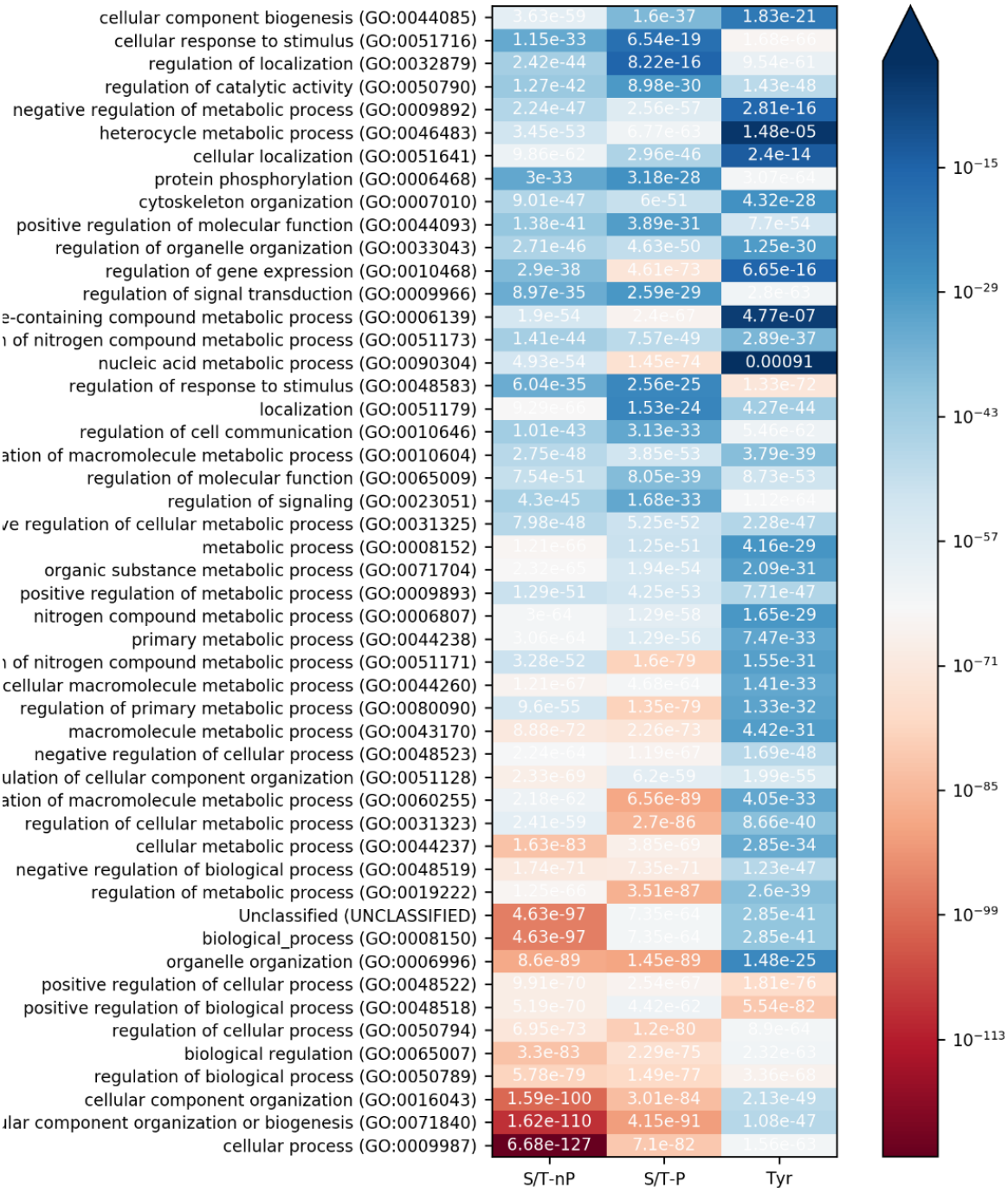

Figure 20. Biological processes associated with specific group of phosphoproteins ('inclusive' scheme)

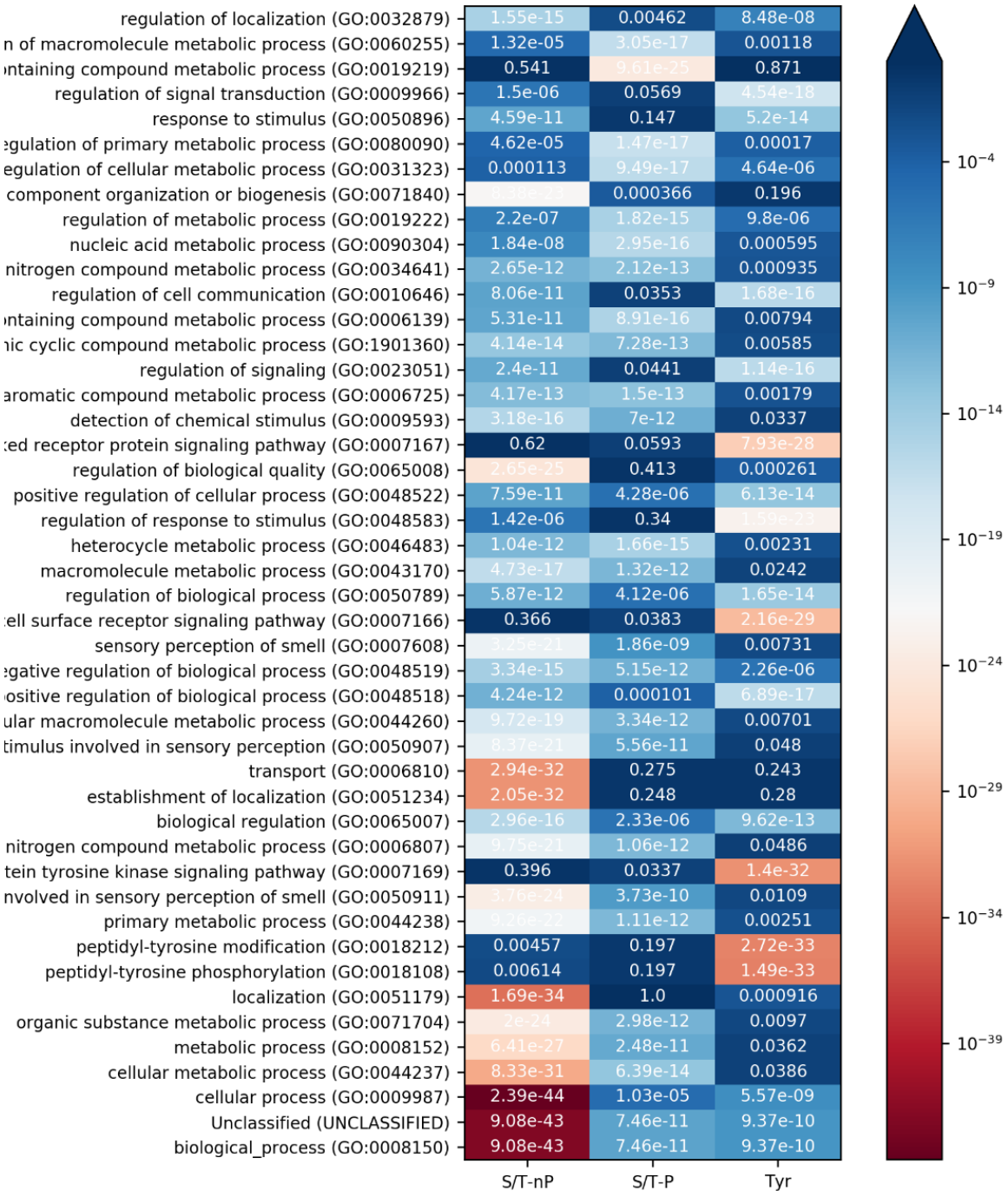

Figure 21. Biological processes associated with specific group of phosphoproteins ('exclusive' scheme)

Other S/T phosphorylation sites were associated with different functionality; protein binding (GO:0005515), olfactory receptor activity (GO:0004984), Ion binding (GO:0043167), small molecule binding (GO:0036094) and carbohydrate derivative binding (GO:0097367) are the examples. These proteins were most likely involved in localization (GO:0051179), transport (GO:0006810), cellular metabolic process (GO:0044237) and regulation of biological quality (GO:0065008).

Proteins with tyrosine phosphorylation sites also showed a strong association with protein tyrosine kinase activity (GO:0004713), transmembrane receptor protein kinase activity (GO:0019199), signaling receptor activity (GO:0038023) and molecular transducer activity (GO:0060089). Biological processes associated with tyrosine phosphorylation sites include peptidyl-tyrosine phosphorylation (GO:0018108), cell surface receptor signaling pathway (GO:0007166), regulation of response to stimulus (GO:0048583) and many terms related to either one of three.

[2-4] Discussion

All these results suggest S/T-P phosphorylation sites could be re-classified as a class which is distinct from other S/T phosphorylation sites and tyrosine phosphorylation sites. S/T-P sites were not only different in local level, including sequence landscape and interacting kinase partner, but also influences phosphoproteins to be associated with different expression patterns, intracellular localizations and biological functions.

It is assumed that both tyrosine kinase family and CMGC kinase family emerged after eukaryotes diverged from other domains of life (126), which might be the reason why proteins with S/T-P sites or tyrosine phosphorylation sites are associated with specific functions and localization patterns. Proteins with S/T-P sites are highly enriched in nucleus and chromatin, which obviously emerged in eukaryotic clade. Also, significant number of those phosphoproteins could bind to nucleic acids as either transcription factors or RNA-binding proteins, which become significantly more diverse in complex organism. Similarly, tyrosine phosphorylation sites are enriched in membrane receptors, and deeply involved in subsequent signaling, another pathway which become substantially intricate in eukaryotes. On the other hand, other S/T

phosphorylation sites were strongly associated with more 'universal' keywords such as cytosol, protein-protein interaction, ion binding and molecular transport. Considering S/T phosphorylation is the ancestral form of protein phosphorylation, I would like to suggest a hypothesis that both S/T-P phosphorylation and tyrosine phosphorylation co-evolved with corresponding kinases to occupy new functional niches emerged in eukaryotic systems.

It is likely that the co-evolution of phosphorylation sites and kinases did not occurred in a single way. For example, plants have tyrosine phosphorylation sites but the kinases responsible for the phosphorylation of tyrosine residues do not fall in classical tyrosine kinase category (127), indicating that the same type of PTM may emerge in completely different contexts independently. However, CMGC kinases are present in every major eukaryotic clades, and regardless of which species it come from, conserved CMGC kinases such as CDKs and MAPKs are largely involved in the same biological process - which suggests the specific niche the S/T-P phosphorylation sites occupy would also be the same overall.

Prokaryotes have no kinases specifically targeting S/T-P sites, which is reflected in the lower frequency of S/T-P sites within phosphoproteome. This suggests the advent of kinases with a preference towards +1 proline was the pivotal event which allowed further differentiation of S/T-P sites. It is likely that the original CMGC kinase gradually acquired preference towards +1 proline: the ancestral CMGC kinase reconstructed from animal and fungal sequences showed preference towards proline and arginine at +1 site (74). However, it could not be determined whether the other protein kinases developed some sort of avoidance mechanism towards +1 proline, or just ignore it.

It is notable that the S/T-P phosphorylation generally lacks positive marker which distinguishes it from other protein sequences. There was no immediately recognizable sequence marker which distinguishes S/T-P sites from non-phosphorylated SP / TP, while overall decreased frequency of aliphatic residues and aromatic residues are observed (Figures 7, 8). Also, CMGC kinases also rely on negative markers for recognition. Most of protein kinases have a specialized structure to recognize specific sequence marker: for example, tyrosine kinases have conserved tryptophan residue which establish pi-pi interaction with the aromatic ring of would-be phosphorylated tyrosine, thereby distinguishing proper substrates from serine/threonine

phosphorylation sites (128). On the other hand, CMGC kinases use an absence of hydrogen bond acceptor to screen out residues other than proline at +1 site. For this reason, some structurally relaxed CMGC kinases may interact with non S/T-P substrates and catalyze phosphorylation (73). While there are several sequence motifs which are recognized by certain members of CMGC kinase family, these motifs are often nebulous and cannot be generalized.

However, there are several biophysical properties which are known to be associated with S/T-P phosphorylation sites, including local conformational changes and peptide extension (102) suggesting the information embedded in the phosphorylation site sequence could manifest into observable features. This raises a hypothesis that the biophysical properties of S/T-P sites are the positive markers of distinguishing S/T-P sites from either other S/T phosphorylation sites or non-phosphorylated SP / TP dipeptides - which is examined in depth in the next chapter.

[3] Phosphorylation of S/T-P sites has different biophysical properties, which are responsible for different consequences after phosphorylation

[3-1] Introduction

In this chapter, I'd like to discuss about biophysical properties of S/T-P sites which not only distinguish those sites from non-phosphorylated sequences and other phosphorylation classes, but also affects the thermodynamics of substrates before and after phosphorylation.

One of the conclusions drawn from the previous chapter was that no viable positive sequence marker other than +1 proline is present for S/T-P sites. However, existence of +1 proline is insufficient to indicate the accompanying serine / threonine is phosphorylated: in fact, only about 10% of SP / TP dipeptides are acknowledged as valid phosphorylation sites (Table 3). From this result, we hypothesized the contribution of hidden information, which could not be identified with simple sequence analysis, would be more pronounced in S/T-P sites.

Groundbreaking work by Dunker and colleagues (102) first associated protein phosphorylation with surrounding intrinsic disorder, and consideration of intrinsic disorder resulted in a phosphorylation site predictor with better predictive performances. Also, similar contribution of conformational thermodynamics were shown by Elam (107) to involve polyproline II (PII) helix propensity of protein sequence around phosphorylation site. These observations suggest a distinct role for the local conformational equilibrium of the phosphorylation site candidates, which might determine not only the activity of the phosphoprotein itself but possibly also the kinase specificity. Also, unique inherent nature of proline (section 1-4) and different sequence environment (section 2-3-1) suggest S/T-P sites might be better discerned by utilizing this information.

To test this hypothesis, I calculated biological properties of phosphorylation sites and non-phosphorylated counterparts and identified properties which are either informative for all phosphorylation sites or only meaningful in identifying specific phosphorylation class. Also, by calculating end-to-end distance change and folded state free energy change caused by phosphorylation, I observed class-specific biophysical

consequences induced by phosphorylation. In addition, different patterns of phosphorylation sites were identified, which suggest that additional differences associated with multiple phosphorylation event might exist between S/T-P sites and other phosphorylation sites.

[3-2] Approaches

-3.2.1. Vertical & horizontal information

Our lab has developed a statistical thermodynamic framework that considers contributions to kinase selectivity driven by two categories of biophysical properties, namely 'vertical' and 'horizontal' information. Vertical information refers to the site-specific properties, requires to be presented at a particular sequence position for recognition. On the other hand, horizontal information refers to ensemble-averaged properties, which is typically not only determined by single amino acid at certain site but conserved along a sequence stretch. For this reason, horizontal information could be conserved among ortholog proteins with divergent sequences (see section 3-2-4).

-3.2.1.1. Vertical information: charge, aromaticity and others

Amino acids at specific positions are often crucial for protein-protein interaction. Binding interface or active site of amino acid commonly form a local structure which only binds to specific amino acids or amino acids with common biophysical properties. For this reason, vertical information has been recognized as a crucial feature of protein-protein interaction.

Charge – both positive and negative – is the most commonly known vertical information in protein. Aspartate, glutamate, lysine and arginine fall in this category: histidine is often mentioned as a positively charged amino acid but its pKa2 fall around 6.0, which means it is usually not charged in cytoplasm, provided the side chain is exposed to solvent (129). These amino acids are particularly important as these could establish ionic bonds or salt bridges with interaction partners and decrease free energy (130). It provides an effective mechanism

of discerning proper substrates from others as electrostatic interaction is not only strong among non-covalent interactions but also selective as it is only formed between opposite charges.

Aromatic amino acids are often required at interaction surfaces. Tyrosine and tryptophan is particularly important: phenylalanine is hydrophobic enough to be buried inside the protein and seldom involved in protein-protein interaction. Non-covalent attraction of aromatic rings (pi stacking) and interaction between aromatic ring and positively charged side chains (cation-pi interaction) allows aromatic amino acids to function as both - a sensor which recognizes specific side chains are present at given location and a signal which could be recognized by a corresponding receptor (131).

Structure of binding site often requires specific side chain property, which is possessed by only a single amino acid type, for stable binding. We already mentioned the example of proline at section 1-3-2, but other amino acids such as cysteine and histidine could be other examples. Cysteines could spontaneously form disulfide bond when two cysteine side chains are near of each other and environmental pH is low. While mostly reversible, disulfide bond is a covalent bond, which is much stronger than other non-covalent interactions and could be sustained for an extended time, allowing binding partners to establish stable complex (132). Histidine is most often associated with metal binding, but its unique imidazole ring structure is also targeted by proteins which specifically recognize it (133). These amino acids contribute to establish substrate specificity along with charge and aromaticity features.

-3.2.1.2. Hydrophobicity

Hydrophobicity is the property of molecules which avoids contact with water molecule. Actually, no repulsive force is involved in this phenomenon; instead, the molecule is simply excluded from hydrogen bond network, which is spontaneously formed in water-based solution to decrease the free energy, and subsequently separated from the solution, which is often perceived as hydrophobic molecules attract each other (hydrophobic effect) (134). In protein science, hydrophobicity generally refers to the tendency of avoiding contact with water of each amino acid types (135). As water is dominant solvent of physiological

systems, amino acids with high hydrophobicity tend to avoid surface exposure and become concentrated in cores of globular proteins.

Hydrophobicity is correlated with different structural properties, such as polarity, accessible surface area (ASA) and contact number. ASA is particularly important in biophysics as it is associated with the thermodynamic properties of protein: it could be utilized to estimate both enthalpy and entropy, which in turn are determinants of free energy (136). IDRs are typically associated with low hydrophobicity, which allows them to be exposed to surface and exert functions based on its structural flexibility. Similarly, PTM sites generally show low hydrophobicity and exposed to the surface, which allows subsequent recognition by corresponding enzymes (137).

-3.2.1.3. Secondary structures

Secondary structure of protein is a three-dimensional conformation, or 'folding pattern' of local sequence within a protein. Amide hydrogens and carbonyl oxygens in a protein backbone form a hydrogen bond to decrease free energy and (consequently stabilize). Depending on the side chains and local thermodynamic environment, different hydrogen binding pattern may arise, which leads to different folding patterns.

Hydrogen bond formation between amino acids is limited by the structural properties of protein backbone. Partial double bond nature of peptide bonds and homochirality of amino acids (except glycine) allow only a small fraction of possible dihedral angle (φ/ψ) pairs of amino acids to be energetically 'allowed' (138). These dihedral angle pairs roughly correlate with known secondary structures: (-57°/-47°) with alpha-helix, (-139°/+135°) with antiparallel beta sheet, (-119°/+113°) with parallel beta sheet, (-49°/-26°) with 3_{10} helix, (+57°/+47°) with left-handed alpha helix, and so on (139). In some cases, inherent properties of amino acid places additional constraint in conformation. Proline, for example, has no amide hydrogen: the side chain is connected to backbone nitrogen and form a pyrrolidine ling structure. Therefore, prolines could have phi angles between -110°~-30° only, and disrupts most beta-sheet conformations (section 1.4).

Secondary structure propensity is a preference of amino acids of adopting specific secondary structure in

physiological condition. These have been one of the most basic means of estimating secondary structure of given protein sequence. These parameters provided a framework for more complex prediction algorithms, starting from Chou-Fasman method in 1970s to sophisticated artificial neural-network (ANN) models (140). Also, propensity scales are often utilized as easy-to-calculate proxies for actual structural information in the prediction of different behaviors, such as DNA binding.

-3.2.1.4. Polyproline II conformation (PII)

Polyproline II conformation (PII) is a special type of secondary structure first identified in polyproline peptide. Polyproline peptide could adopt either of two distinct conformations depending on its stereochemistry, and PII is formed by trans- isomers of peptide bonds (141).

PII is a left-handed helical structure with dihedral angles of (-75°/+145°), which the values are superficially close to those of beta-sheet conformation, and three residues consist of one turn. PII is highly extended (or relaxed) conformation with translation, vertical distance between amino acids, of 3.12 angstroms; for alpha helix, translation is just 1.5 angstroms. It is notable that no hydrogen bond is formed internally: as opposed to other helical conformations, carbonyl oxygen and (if exists) amide hydrogen are exposed to surface and form hydrogen bond with other molecules to stabilize energetically. For this reason, PII region is often involved in protein-protein interactions or interdomain interactions. SH2/3, WW, GYF, UEV and EVH1 are examples which interact with PII conformation (142).

This conformation is now understood as a characteristic of not only proline-rich regions but also a wider range of IDRs. For example, CD spectra (which is different from random coils) typical to PII was similarly observed in polyglutamate and polylysine (143), suggesting PII conformation could form in denatured polypeptides containing minimal or no prolines. On the other hand, PII makes up about 2% of structures found in protein data bank (PDB), and hundreds of analyzed binding regions adopt PII conformation in physiological condition.

Also, there is evidence which supports preference of phosphorylation sites towards high PII. Phosphorylation

is the most statically enriched feature in high PII proteins in the human proteome (107). Interestingly, PII propensity change caused by phosphorylation is highly dependent on the base amino acid types: while phosphoserine has similar PII propensity to serine (0.24 → 0.26), phosphothreonine has significantly higher PII propensity (0.30 → 0.92) and phosphotyrosine has one of the lowest PII propensity value (0.14 → 0.09).

PII propensity scale is a quantitative measure of preference towards PII conformation, just as propensity scales for other secondary structures such as alpha helix and beta sheet. Several independent scales exist, which utilizes different experimental approaches to assign values for amino acid types; in this study, scale devised from my lab using isothermal titration calorimetry (ITC) is used (107). This scale was found to fit the best with values independently calculated using molecular dynamics simulation (144).

-3.2.2. COREX/eSCAPE

Another set of parameters utilized was imported from COREX/eSCAPE (or simply eSCAPE), an algorithm which predicts base thermodynamics of given protein solely based on amino acid sequence (145). eSCAPE is derived from COREX, which models conformational ensemble and its energetic properties of protein from its three-dimensional structure (146).

COREX algorithm is divided into three steps: enumeration of ensemble, calculation of relative free energy for each microstate, and characterization of ensemble thermodynamics. First step involved partitioning of protein sequence into smaller folding units, with window size of ~10. It is assumed that each folding unit is either completely folded (native state) or completely unfolded: this produces $M*(2^N-2)$ partially native states, M for window size and N for number of folding units, which represents every combinations possible by partitioning.

Second step involved calculation of statistical weight of each conformational state. Statistical weight is calculated from free energy state, which is in turn calculated with Gibbs-Helmholtz equation.

$$K_i = e^{=\Delta G_i/RT} \quad \ldots (2)$$

$$\Delta G_i(T) = \Delta H_i(T_{ref}) - T\Delta S_i(T_{ref}) + \Delta Cp_i[(T - T_{ref}) - \mathrm{Tln}(\mathrm{T}/T_{ref})] \quad \dots (3)$$

Here, heat capacity and enthalpy were approximated from apolar and polar solvent-accessible surface areas, which could be determined from input three-dimensional structure. Reference temperature was 60℃.

$$\Delta\mathrm{H}(60) = \Delta H_{ap}(60) + \Delta H_{pol}(60) = -8.44ASA_{ap} + 31.4ASA_{pol} \quad \dots (4)$$

$$\Delta\mathrm{Cp} = \Delta Cp_{ap} + \Delta Cp_{pol} = 0.45\Delta ASA_{ap} - 0.26\Delta ASA_{pol} \quad \dots (5)$$

On the other hand, entropy term is decomposed into solvent entropy and conformational entropy. Solvent entropy could be calculated from apolar and polar heat capacity contributions as below.

$$\Delta S_{total} = \Delta S_{solv} + \Delta S_{conf} \quad \dots (6)$$

$$\Delta S_{solv,Tot}(T) = \Delta Cp_{ap}\ln(\mathrm{T}/385) - \Delta Cp_{pol}\ln(\mathrm{T}/385) \quad \dots (7)$$

Conformational entropy consists of three terms; entropy change for originally buried side chains which become exposed in a microstate; entropy change for originally exposed side chains upon unfolding of peptide backbone; and entropy change for backbone which become unfolded in a microstate. Individual values for each amino acid type are imported from previous studies.

$$\Delta S_{conf} = \Delta S_{bu-ex,i} + \Delta S_{ex-un,i} + \Delta S_{bb,i} \quad \dots (8)$$

Probability of each conformational state is calculated by dividing the statistical weight of given state by the sum of weights of all possible states.

$$P_i = \frac{K_i}{\sum_{i=1}^{N\,states} K_i} \quad \dots (9)$$

From here, we could calculate residue stability constant for each amino acids. This gives an estimate about how the given amino acid would behave in conformational ensemble.

$$\kappa_{f,j} = \frac{\sum P_{f,j}}{\sum P_{nf,j}} \quad \dots (10)$$

COREX enables estimation of thermodynamic behaviors of protein with known structure with reasonable

accuracy, but there are two inherent issues. First, COREX is only applicable when there is a known structure for target protein, which is not true for the most IDR sequences. Second, calculation of $M*(2^N-2)$ possible states make full COREX calculation computationally intensive, making analysis of thousands of protein sequences not so feasible.

eSCAPE is a further simplification of COREX by implementing machine learning approach and thereby removing three-dimensional requirement. It starts from the assumption that thermodynamic properties for given position is affected by adjacent amino acids, making tripeptide an informational unit. From 122 X-ray crystallography data, triplet patterns of free energy, apolar and polar enthalpy, and entropy values were collected and parametrized. Then, linear regression model was applied to calculate the thermodynamic descriptors for specific sites as follows:

$$\Delta G = \left(\left(0.8195 \times min_{i,\Delta G}\right) + \left(0.7492 \times max_{i,\Delta G}\right)\right) + 4696 \quad \ldots (11)$$

$$\Delta H_{ap} = \left(\left(0.7665 \times min_{i,\Delta H_{ap}}\right) + \left(0.7632 \times max_{i,\Delta H_{ap}}\right)\right) - 5068 \quad \ldots (12)$$

$$\Delta H_{pol} = \left(\left(0.791 \times min_{i,\Delta H_{po1}}\right) + \left(0.7524 \times max_{i,\Delta H_{pol}}\right)\right) + 6195 \quad \ldots (13)$$

$$T\Delta S = \left(\left(0.7047 \times min_{i,T\Delta S}\right) + \left(0.7507 \times max_{i,T\Delta S}\right)\right) + 1998 \quad \ldots (14)$$

This allowed calculation of needed thermodynamic features within reasonable time period. I used free energy, apolar and polar enthalpy, and entropy values for both folded (native) and unfolded (denatured) states for the analysis of phosphorylation sites and formulation of phosphorylation site predictors.

-3.2.3. Hydrodynamic radius / end-to-end distance

Hydrodynamic radius and end-to-end distance refer to an 'effective size' of molecule in terms of fluid dynamics. It provides a rough idea about the compactness and the shape of given molecule: for example, even if the molecular weights are the same, hydrodynamic radius of denatured protein is typically much

larger than that of folded protein. Hydrodynamic radius and end-to-end distance could be converted to each other by using this equation 15:

$$\vec{R} = R_h \cdot \sqrt{6} \quad \ldots (15)$$

While IDRs are often approximated with random coils, the actual behaviors of those are often significantly different largely due to two biophysical properties: charge and PII propensity. Interaction between charged amino acids is strong enough to disrupt the assumption that there is no significant interaction between distal side chains. PII propensity, on the other hand, promotes formation of PII conformation with long end-to-end distance, leading to longer hydrodynamic radius overall. By merging equations from previous studies, it was able to approximate hydrodynamic radius from charge and PII propensity (144, 147). This power-law equation could be used to calculate both global hydrodynamic radius of entire protein and local hydrodynamic radius which describes persistence length of short peptide regions.

$$R_h = R_0 \cdot N^{\nu} \quad \ldots (16)$$

$$\nu(f_{PII}, |Q|) = \nu_0 + \alpha \cdot s(|Q|) + \beta \cdot (1 - s(|Q|)) \cdot \ln(1 - f_{PII}) \quad \ldots (17)$$

Here, R_h is hydrodynamic radius of peptide, R_0 is hydrodynamic radius of single amino acid (which is 2.16Å), N is length of peptide, f_{PII} is PII propensity of peptide, s(|Q|) is a sigmoid function fitted with net charge and hydrodynamic radius (147), α and β are scaling coefficients for the effects of net charge and PII propensity respectively. This equation suggests increase of both net charge and PII propensity increases hydrodynamic radius, albeit in different degrees.

Local hydrodynamic radius could be important for phosphorylation sites for two different reasons. First, most of known structures of kinases have a groove-like active site which binds to extended substrates (123): peptides with other secondary structures, such as alpha helix, could not fit in this groove and consequently excluded from phosphorylation targets. While the specific three-dimensional arrangements of side chains within active sites, which largely determine more detailed substrate preference, are different between individual kinases, the overall architecture of protein kinases are originated from the common ancestor and therefore largely shared between each other. From this, we could expect extended protein conformation to be

a general requirement for kinase binding.

On the other hand, phosphorylation not only adds double negative charge but also causes residue-specific changes in PII propensity, which in turn changes local hydrodynamic radius and bring number of consequences (148). In particular, phosphorylation of threonine brings the biggest change in PII propensity and therefore is expected to induce the biggest changes in conformational dynamics. In contrast, phosphorylation of tyrosine slightly decreases PII propensity, which likely counterbalances the effects of double negative charge and minimize the consequences.

-3.2.4. Conservation of vertical & horizontal information

To visualize the conservation of vertical and horizontal information, I utilized the ortholog sequences of human proteins with DNA-binding transcription factor activity (GO: 0003700). Transcription factors are likely to have both structured and intrinsically disordered regions (149), which enables analyzing different degree of evolutionary conservation within a single protein. We selected ortholog groups with the number of members between $10 < n < 250$ in OMA database (150) and downloaded multiple sequence alignments as archived in the database.

Normalized sequence conservation scores for the local sequence alignment were calculated as below. The multiple sequence alignment with size =n was divided into windows (size = 5) which overlaps with each other. For each window, pairwise local alignment scores using BLOSUM62 matrix (151) were calculated between a reference sequence (Seq_i) and each of all other sequences within same ortholog group (Seq_j). This step was repeated using each of the ortholog sequences in the alignment as a reference. For each iteration, every pairwise alignment scores were divided by the maximum score attainable: the case when a sequence which is identical to the reference was applied for comparison. Calculated pairwise scores were averaged to obtain a normalized local sequence conservation score as in equation 18:

$$Score_{seq} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n\,(not\ i)} \frac{BLOSUM\,(seq_i, seq_j)}{BLOSUM\,(seq_i, seq_i)}}{n(n-1)} \quad \ldots (18)$$

Native state free energy for each protein sequences was calculated using the eSCAPE algorithm (146). For the same windows we utilized for the calculation of local sequence conservation scores, we obtained local average along with standard deviation. Horizontal conservation score was computed using the following equation 19:

$$Score_{Hor} = 1 - \frac{SD_{local}}{C_s} \quad \ldots (19)$$

In this case, scaling coefficient (C_s= 3.3 (kcal/mol)) was calculated from 10 different ortholog groups exhibiting high sequence conservation and structural stability (for example, actin (ACTB) and rhodopsin (RHO) families). To observe the correlation with free energy, BLOSUM-based sequence conservation scores and horizontal conservation scores were normalized again with μ = 0 and SD = 1 (i.e. a Z-score). Linear correlations between average free energy and both conservation scores were calculated subsequently (see Figure 53).

-3.2.5. Data sources

The same datasets for phosphorylation sites and non-phosphorylated sequences used in chapter 2 were again utilized for the analysis. 544 amino acid-based scales were retrieved from AAindex database (152). PII propensity scale for canonical and phosphorylated residues were experimentally measured by Elam (107). Disprot intrinsically disordered protein scale was retrieved from this paper (153). COREX/eSCAPE thermodynamic predictor was created by Gu (146). Experimentally identified IDR annotations and sequences were retrieved from Disprot database (154).

[3-3] Results

-3.3.1. S/T-P sites have distinct biophysical fingerprints

Calculation of biophysical properties around phosphorylation sites and non-phosphorylated sequences not only confirmed prior knowledge or findings discussed in the previous chapter, but also revealed features which is specifically associated with a particular class of phosphorylation sites.

Phosphorylation sites were found to be biased towards high polarity (Figure 22), interactivity, flexibility and surface exposure scales. In contrast, hydrophobicity (Figure 23), local stability and buried propensity values were substantially low around phosphorylation sites. These results were consistent with our expectation, as non-spontaneous PTM sites in general should be exposed to the surface to interact with enzymes and subsequently get modified.

Among secondary structure-associated terms, beta-sheet propensity values were found to be significantly lower around phosphorylation sites (Figure 24). It is not clear whether this result implies genuine avoidance of beta-sheet conformation by phosphorylation sites, as beta-sheet propensity scales in AAindex database were invariably correlated with hydrophobicity scales, leaving the possibility of being a duplicate result. Besides, alpha helix propensity values were, except for a particular case that would be discussed later, not found to be meaningfully different between phosphorylation sites and non-phosphorylated sequence (Figure 25).

There were biophysical properties which were associated only with specific classes of phosphorylation sites (Figure 26). Distribution of charges (Figure 27) was informative for discerning S/T-nP sites and tyrosine phosphorylation sites from non-phosphorylated counterparts but not for S/T-P sites. The information content mainly came from distribution of negative charges (Figure 28) while positive charges (Figure 29) contributed less. While the higher contribution of negative charges was somewhat unexpected, these results were also consistent with our understanding about roles of neighboring charged residues in kinase-substrate recognition (Tables 4, 5).

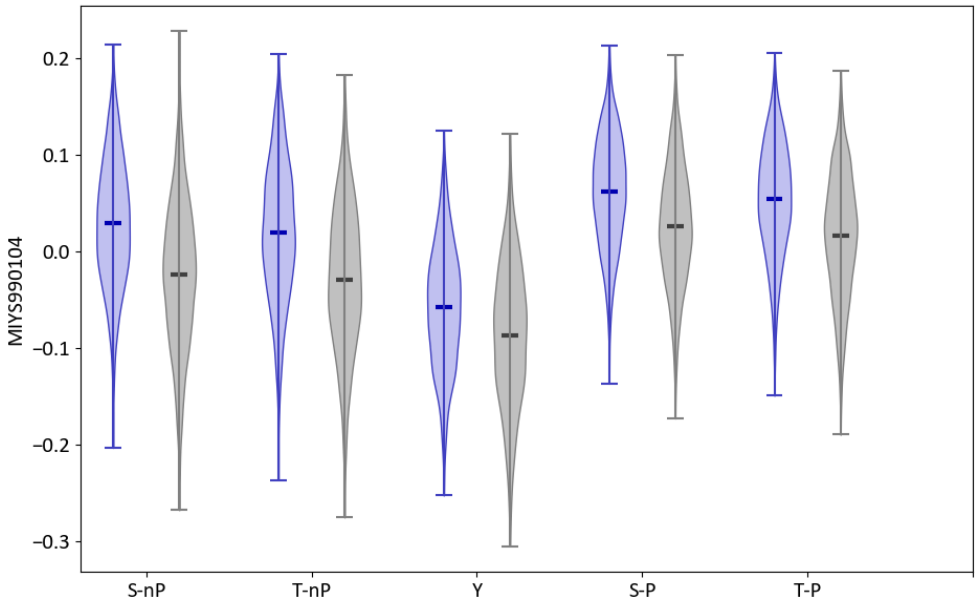

Figure 22. Polarity values (MIYS990104) calculated for phosphorylated and non-phosphorylated protein sequences. (Blue = phosphorylation sites, Gray = Non-phosphorylated sequences)

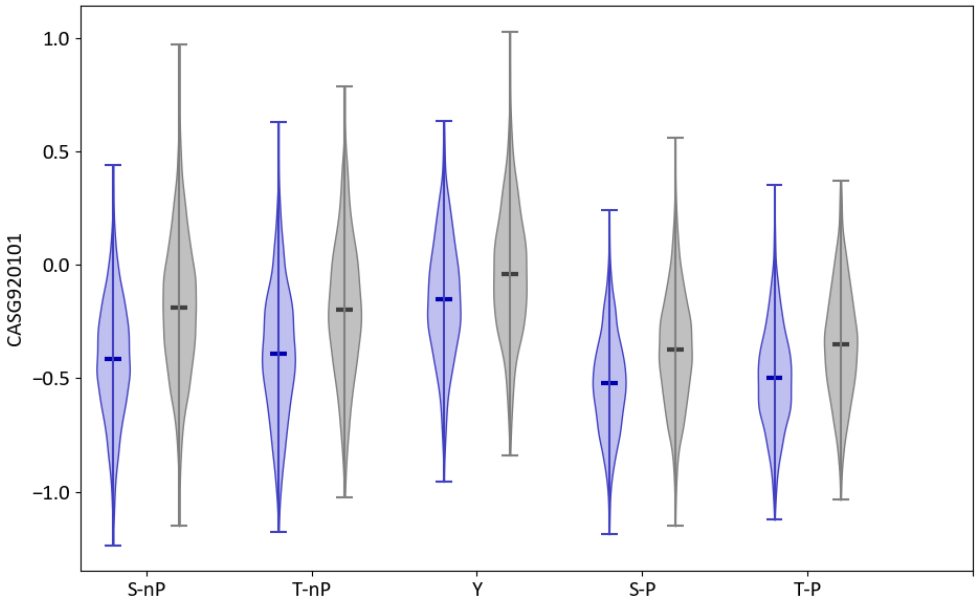

Figure 23. Hydrophobicity values (CASG920101) calculated for phosphorylated and non-phosphorylated protein sequences. (Blue = phosphorylation sites, Gray = Non-phosphorylated sequences)

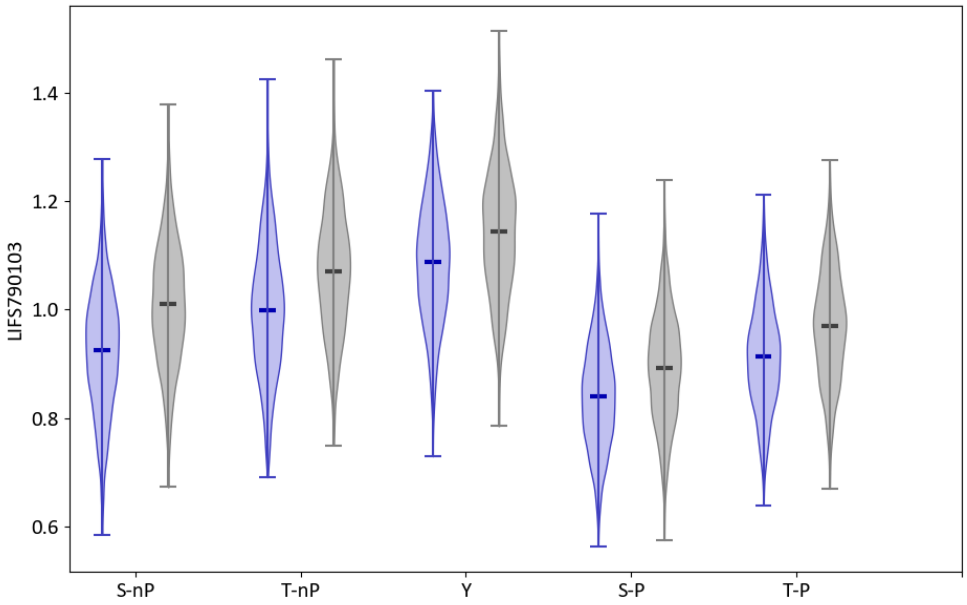

Figure 24. Beta-sheet propensity values (LIFS790103) calculated for phosphorylated and non-phosphorylated protein sequences. (Blue = phosphorylation sites, Gray = Non-phosphorylated sequences)

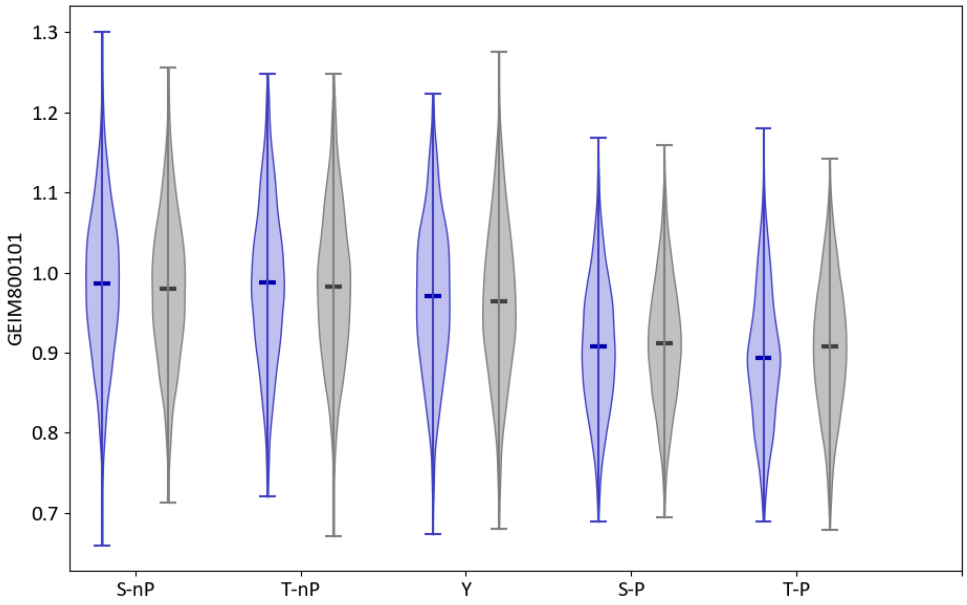

Figure 25. Alpha-helix propensity values (GEIM800101) calculated for phosphorylated and non-phosphorylated protein sequences. (Blue = phosphorylation sites, Gray = Non-phosphorylated sequences)

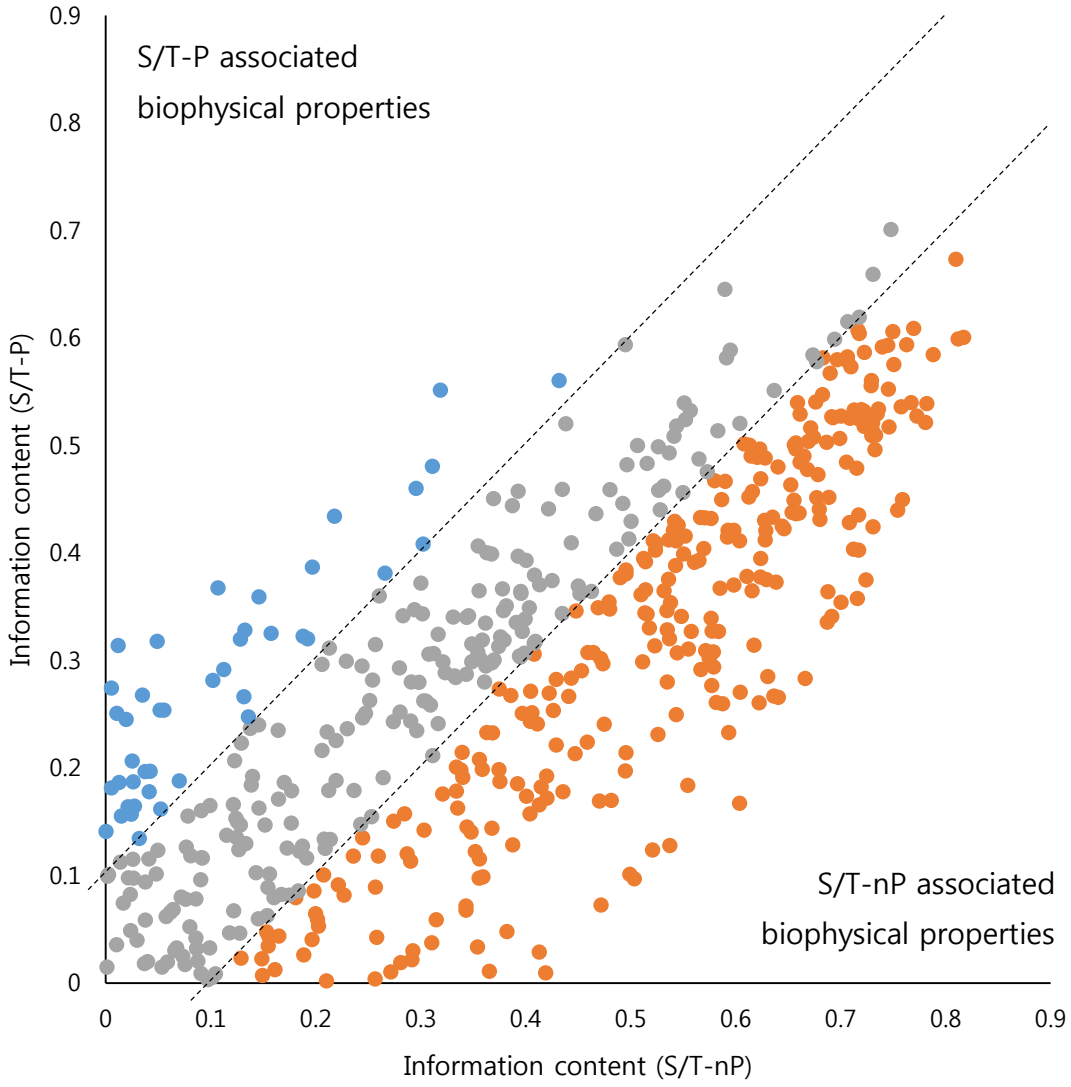

Figure 26. Biophysical properties convey different amount of information for S/T-nP and S/T-P classes. (Blue: properties which are more informative in prediction of S/T-P sites, Orange: properties which are more informative in prediction of S/T-nP sites)

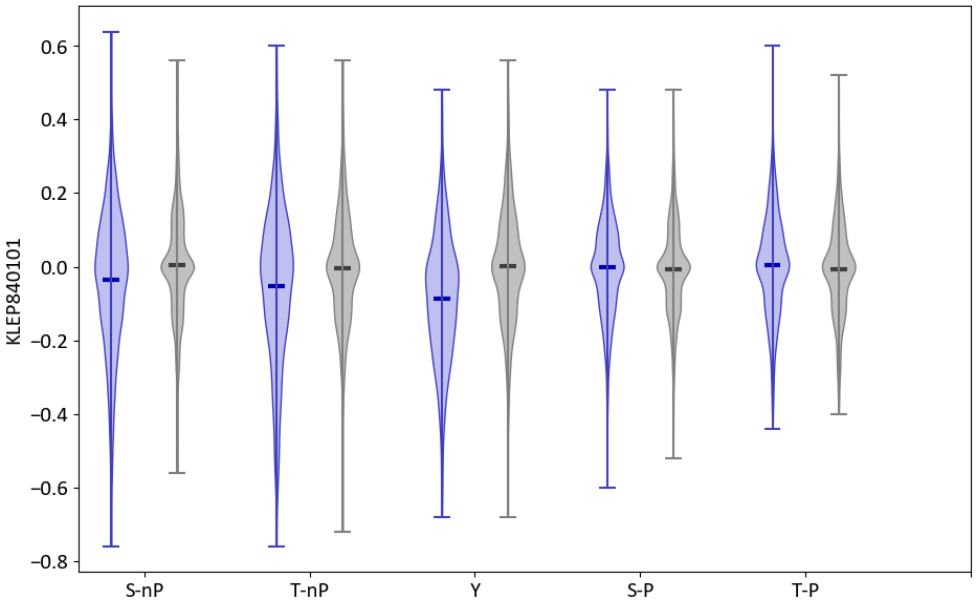

Figure 27. Net charge values (KLEP840101) calculated for phosphorylated and non-phosphorylated protein sequences. (Blue = phosphorylation sites, Gray = Non-phosphorylated sequences)

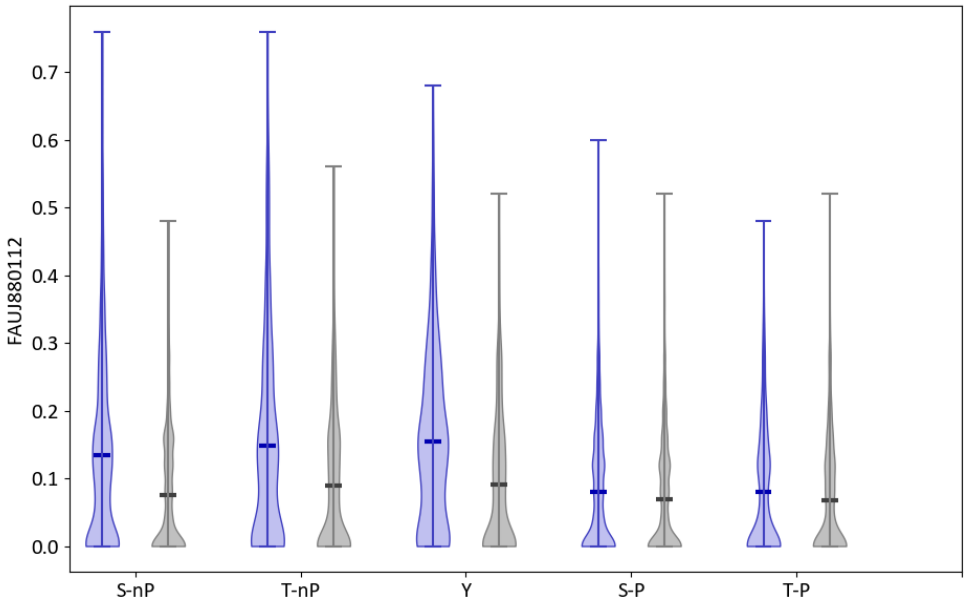

Figure 28. Negative charge contents (FAUJ880112) calculated for phosphorylated and non-phosphorylated protein sequences. (Blue = phosphorylation sites, Gray = Non-phosphorylated sequences)

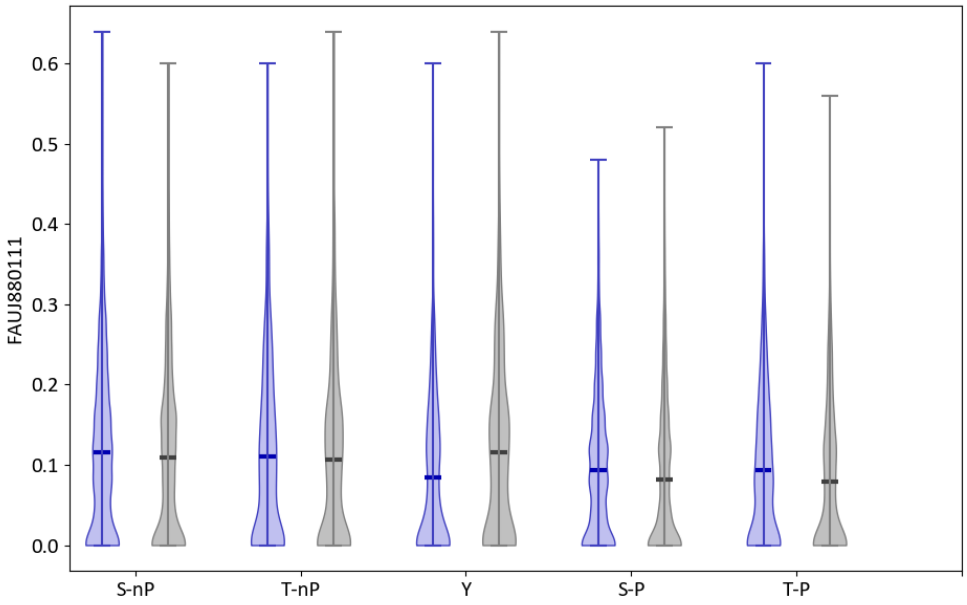

Figure 29. Positive charge contents (FAUJ880111) calculated for phosphorylated and non-phosphorylated protein sequences. (Blue = phosphorylation sites, Gray = Non-phosphorylated sequences)

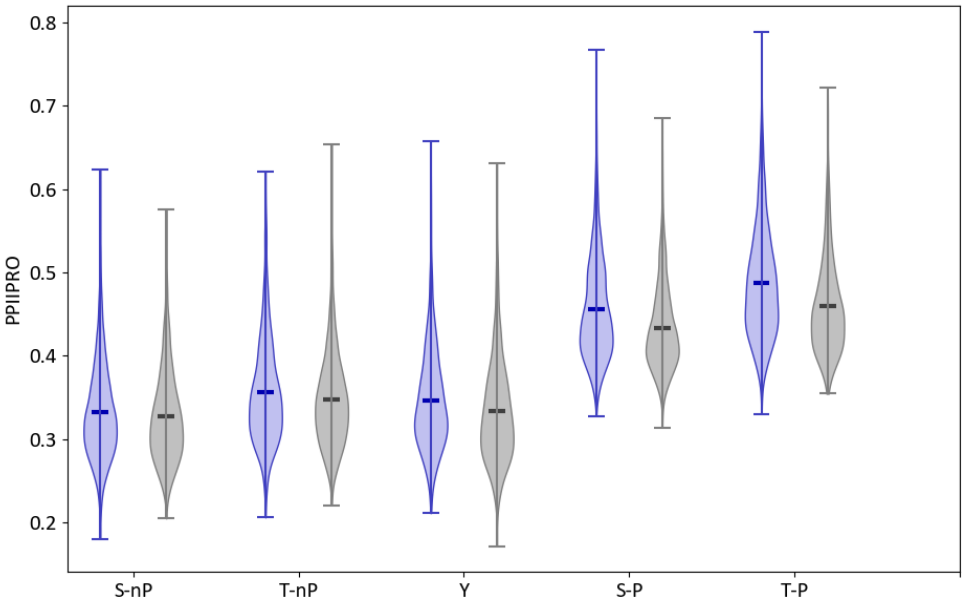

Figure 30. PII propensity calculated for phosphorylated and non-phosphorylated protein sequences. (Blue = phosphorylation sites, Gray = Non-phosphorylated sequences)

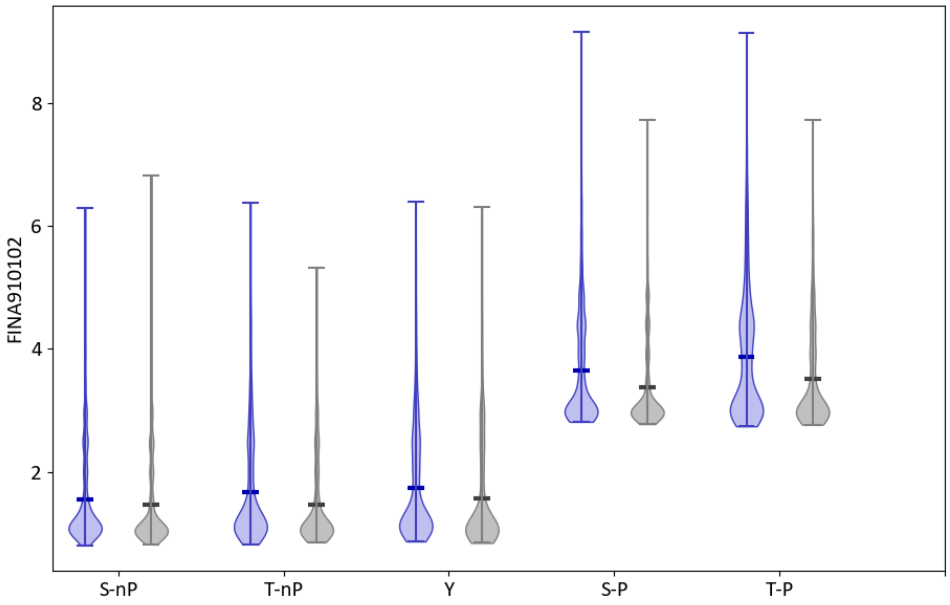

Figure 31. C'-terminal alpha helix propensity values (FINA910102) calculated for phosphorylated and non-phosphorylated protein sequences. (Blue = phosphorylation sites, Gray = Non-phosphorylated sequences)

On the other hand, S/T-P sites had significantly higher PII propensity (Figure 30) and C'-terminal alpha helix propensity (Figure 31) than its non-phosphorylated counterparts. These patterns of differences were not observed for S/T-nP sites and tyrosine phosphorylation sites. PII propensity difference was actually more pronounced between S/T-P sites and S/T-nP sites, which not only suggests the role of PII propensity in distinguishing phosphorylation sites from non-phosphorylated sequences but also strongly implies different local conformational environment around phosphorylation sites, which suggests different consequences after phosphorylation.

-3.3.2. S/T-P sites have higher native state free energy and polar enthalpy

Calculation of thermodynamic environment around phosphorylation sites with COREX/eSCAPE allowed to further elucidate general and class-specific properties of phosphorylation sites.

Phosphorylation sites had ~ 0.42kcal/mol higher native state free energy ($\Delta\Delta G_N$) than non-phosphorylated counterparts in average (Figure 32, Table 7). The difference of free energy was marginally bigger for S/T-P sites (~ 0.55kcal/mol) and lower for S/T-nP sites (~ 0.36kcal/mol). As the free energy denotes the stability of native fold, this result supports previously established relationship between intrinsic disorder of protein and protein phosphorylation (section 1-8) (102).

Investigation of enthalpy and entropy terms followed to identify the origin of differences. Apolar energy difference ($\Delta\Delta H_{ap, N}$) (Figure 33) was significantly lower for tyrosine phosphorylation sites (~ -0.54kcal/mol), followed by S/T-P sites (~ -0.75kcal/mol) and S/T-nP sites (~ -0.98kcal/mol). Considering enthalpy values are inherently tied with accessible surface area (ASA) of peptides, lower apolar enthalpy denotes smaller nonpolar surface exposed to the solvent, thus the result is consistent with the observed relationship between phosphorylation sites and hydrophobicity (Figure 23).

On the other hand, more pronounced differences in polar enthalpy ($\Delta\Delta H_{pol, N}$) (Figure 34) between S/T-P and S/T-nP classes was observed. Polar enthalpy differences between phosphorylation sites and non-phosphorylated sequences were ~ 0.45kcal/mol for S/T-P sites, ~ -0.07kcal/mol for S/T-nP sites and ~0.32

Phospho-site	S-nP	T-nP	Y	S-P	T-P
ΔG_N	-8.20	-7.73	-8.94	-7.14	-6.51
$\Delta H_{ap,N}$	8.93	9.07	10.80	8.58	8.50
$\Delta H_{pol,N}$	-12.26	-11.56	-12.26	-10.75	-9.67
$T\Delta S_{conf,N}$	-4.56	-4.60	-4.72	-4.29	-4.11
ΔG_D	8.66	9.21	9.18	8.61	9.25
$\Delta H_{ap,D}$	-0.78	-0.84	-0.29	-0.90	-1.01
$\Delta H_{pol,D}$	-0.55	0.19	-0.0	-0.33	0.44
$T\Delta S_{conf,D}$	-9.39	-9.42	-9.71	-9.14	-9.06

Nonphos-site	S-nP	T-nP	Y	S-P	T-P
ΔG_N	-8.57	-8.10	-9.43	-7.63	-7.15
$\Delta H_{ap,N}$	9.97	9.93	11.43	9.31	9.33
$\Delta H_{pol,N}$	-12.17	-11.56	-12.58	-11.11	-10.24
$T\Delta S_{conf,N}$	-4.57	-4.60	-4.76	-4.39	-4.26
ΔG_D	8.70	9.32	9.07	8.70	9.19
$\Delta H_{ap,D}$	-0.67	-0.77	-0.24	-0.76	-0.82
$\Delta H_{pol,D}$	-0.25	0.47	0.11	-0.22	0.39
$T\Delta S_{conf,D}$	-9.29	-9.37	-9.55	-9.22	-9.16

P to nP difference	S-nP	T-nP	Y	S-P	T-P
$\Delta\Delta G_N$	0.37	0.37	0.49	0.49	0.64
$\Delta\Delta H_{ap,N}$	-1.04	-0.86	-0.63	-0.74	-0.82
$\Delta\Delta H_{pol,N}$	-0.09	0.01	0.32	0.36	0.57
$\Delta T\Delta S_{conf,N}$	0.02	0.01	0.04	0.10	0.14
$\Delta\Delta G_D$	-0.04	-0.11	0.11	-0.09	0.06
$\Delta\Delta H_{ap,D}$	-0.11	-0.06	-0.05	-0.14	-0.19
$\Delta\Delta H_{pol,D}$	-0.31	-0.28	-0.11	-0.11	0.05
$\Delta T\Delta S_{conf,D}$	-0.10	-0.05	-0.16	0.08	0.10

Table 7. Thermodynamic descriptors calculated with COREX/eSCAPE for phosphorylated / non-phosphorylated sequences (kcal/mol)

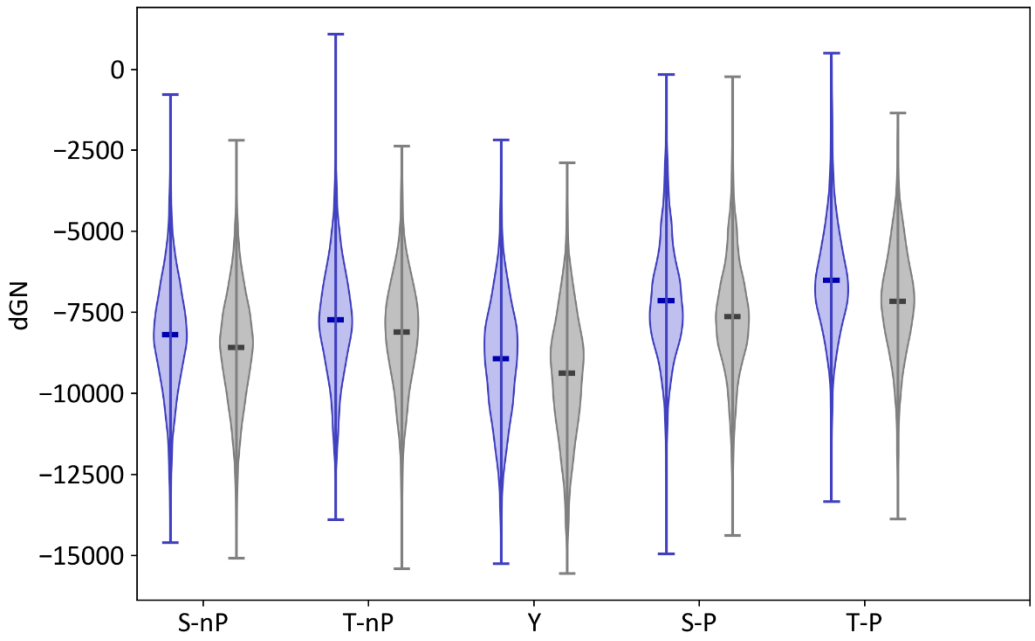

Figure 32. Native state free energy ($\Delta\Delta G_N$) calculated with COREX-eSCAPE for phosphorylated and non-phosphorylated protein sequences (cal/mol)

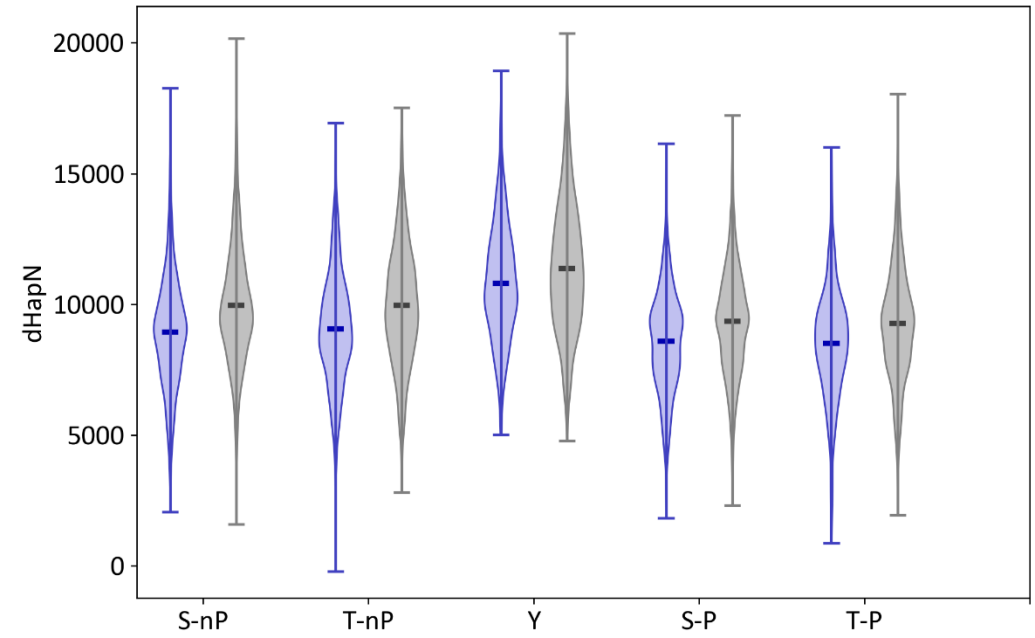

Figure 33. Native state apolar enthalpy ($\Delta\Delta H_{ap,\ N}$) calculated with COREX-eSCAPE for phosphorylated and non-phosphorylated protein sequences (cal/mol)

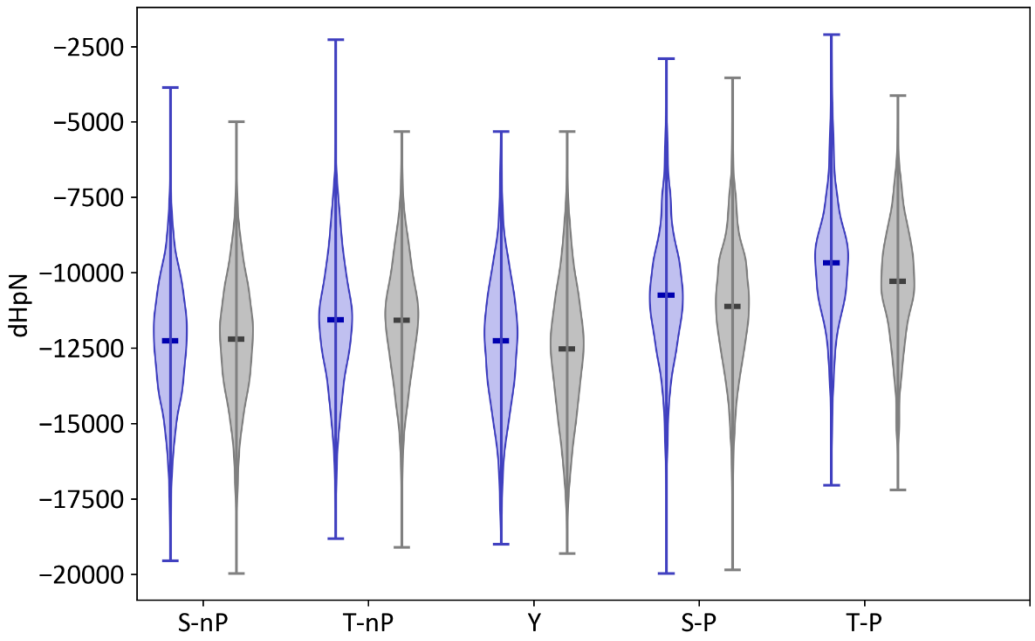

Figure 34. Native state polar enthalpy ($\Delta\Delta H_{pol,\,N}$) calculated with COREX-eSCAPE for phosphorylated and non-phosphorylated protein sequences (cal/mol)

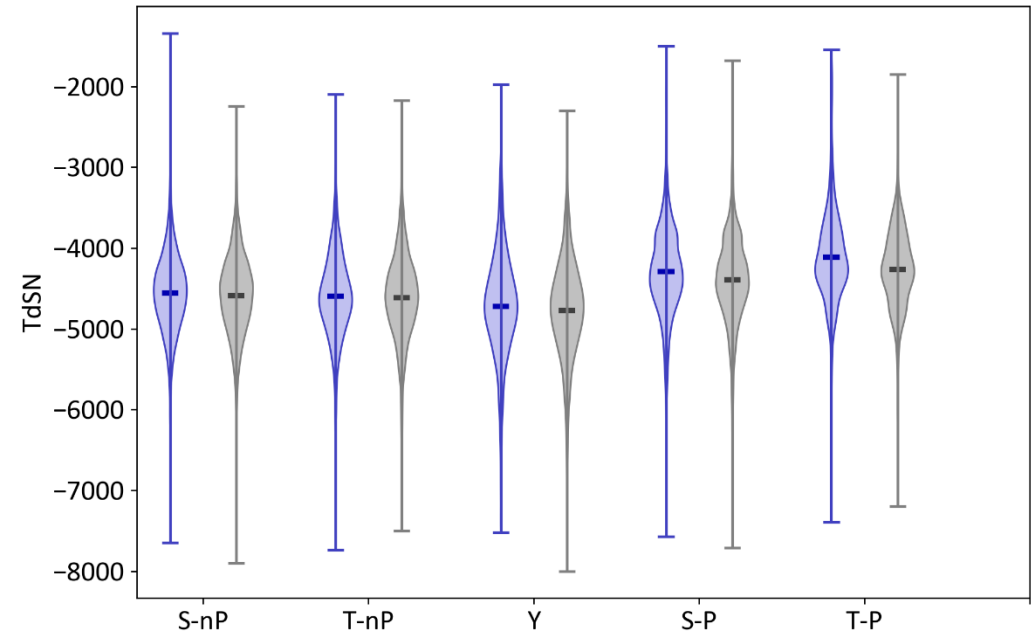

Figure 35. Native state conformational entropy ($\Delta T\Delta S_{conf,\,N}$) calculated with COREX-eSCAPE for phosphorylated and non-phosphorylated protein sequences (cal/mol)

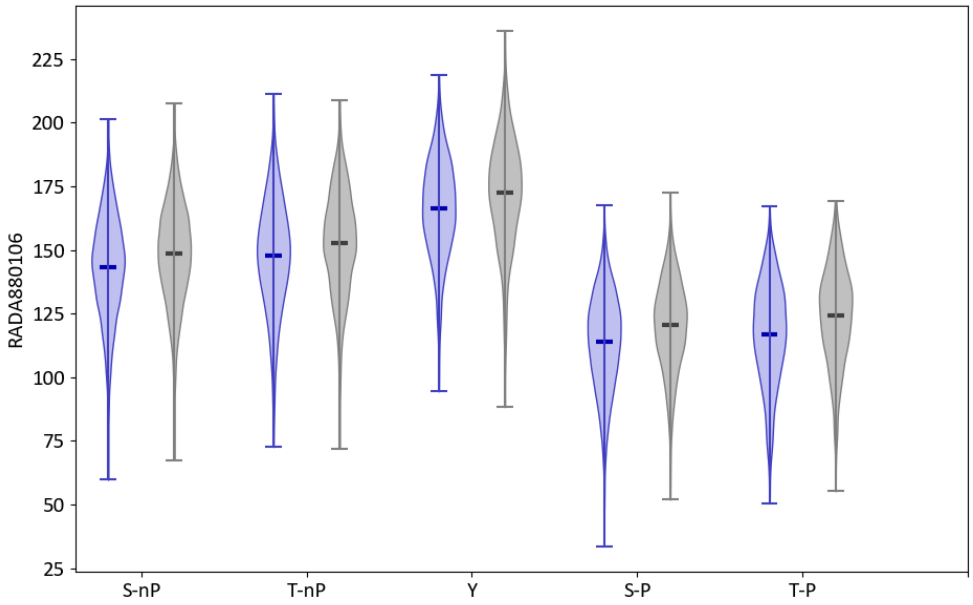

Figure 36. Accessible surface area (RADA880106) calculated for phosphorylated and non-phosphorylated protein sequences

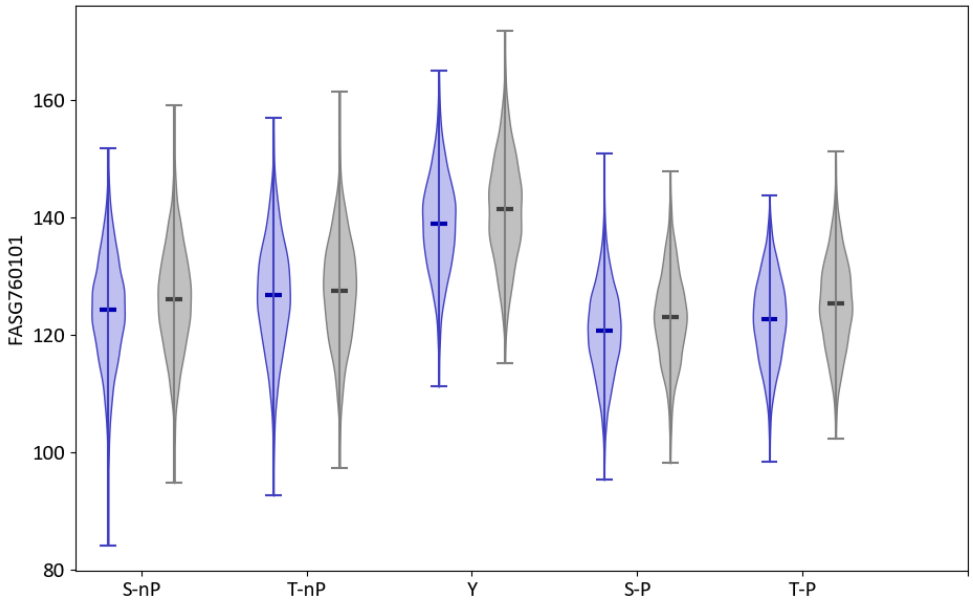

Figure 37. Average molecular weight of amino acids (FASG760101) calculated for phosphorylated and non-phosphorylated protein sequences

kcal/mol for tyrosine phosphorylation sites. This result denotes that S/T-P sites have smaller ASA overall (not just nonpolar ASA): and this result is consistent with independent analysis results which show S/T-P sites are more strongly biased towards lower accessible surface area (RADA880106) and lower molecular weights (FASG760101) (Figures 36, 37).

Similarly, conformational entropy differences ($\Delta T\Delta S_{conf,\ N}$) (Figure 35) between phosphorylation sites and non-phosphorylated sequences were ~ 0.12kcal/mol for S/T-P sites, ~ 0.0kcal/mol for S/T-nP sites and ~ 0.04kcal/mol for tyrosine phosphorylation sites. Thermodynamic descriptors associated with unfolded states were largely uninformative for discerning phosphorylation sites except for unfolded state polar enthalpy difference ($\Delta\Delta Hpol_D$) (Table 7), which was bigger for S/T-nP classes. However, the information associated with descriptors mentioned was not significant enough to construct an independent hypothesis based on these differences.

Focusing on differences between different classes of phosphorylation sites reveals another narrative. Average native state free energy of phosphorylation sites were ~ -7.14kcal/mol for S-P sites, ~ -6.51kcal/mol for T-P sites, ~ -8.20kcal/mol for S-nP sites, ~ -7.73kcal/mol for T-nP sites, and ~ -8.94kcal/mol for tyrosine phosphorylation sites (Table 7). This suggests S/T-P sites are in 1.1~1.2kcal/mol higher free energy environment than S/T-nP sites, further supports its association with IDRs. This difference was largely originated from polar enthalpy, which was 1.5~1.9kcal/mol higher for S/T-P sites, while apolar enthalpy and conformational entropy reduced the difference between classes.

While +1 proline is the biggest contributor to its high free energy value, substitution of proline residue to other random amino acid or vice versa was predicted to be not sufficient to completely remove this free energy difference (Figure 38, Table 8), especially for threonine phosphorylation sites – suggesting the difference between S/T-P sites and S/T-nP sites came from not only +1 residue but also amino acids in vicinity of phosphorylated residues.

Simulating thermodynamic environment of phosphorylated sequences by serine (S) to aspartate (D) or threonine (T) to glutamate (E) produced interesting results (Figure 39, Table 9). While the phosphomimetic

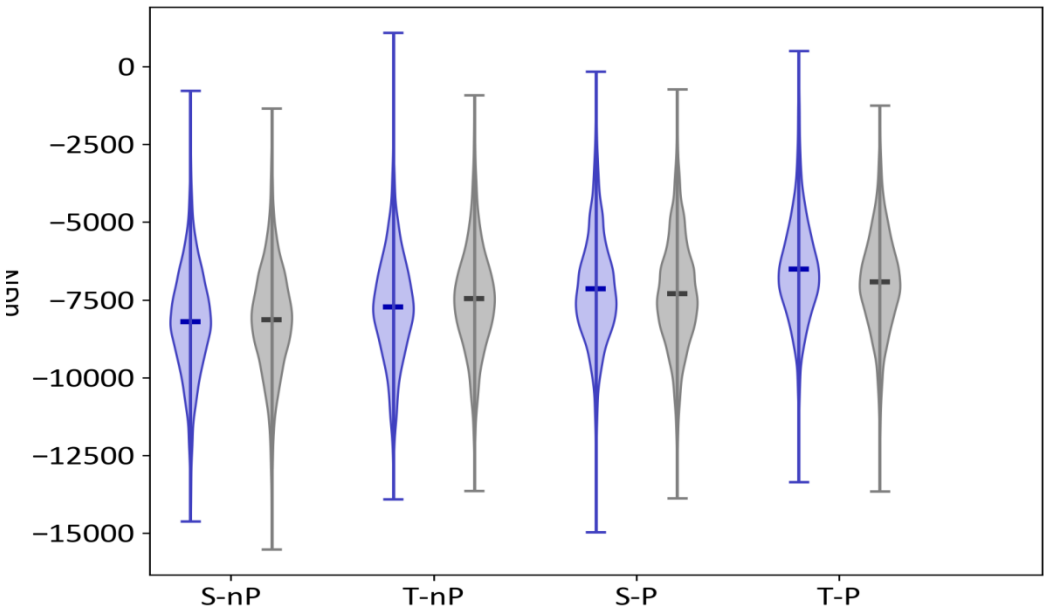

Figure 38. Native state free energy change predicted for phosphorylation sites with +1 site substitution (P ↔ nP)

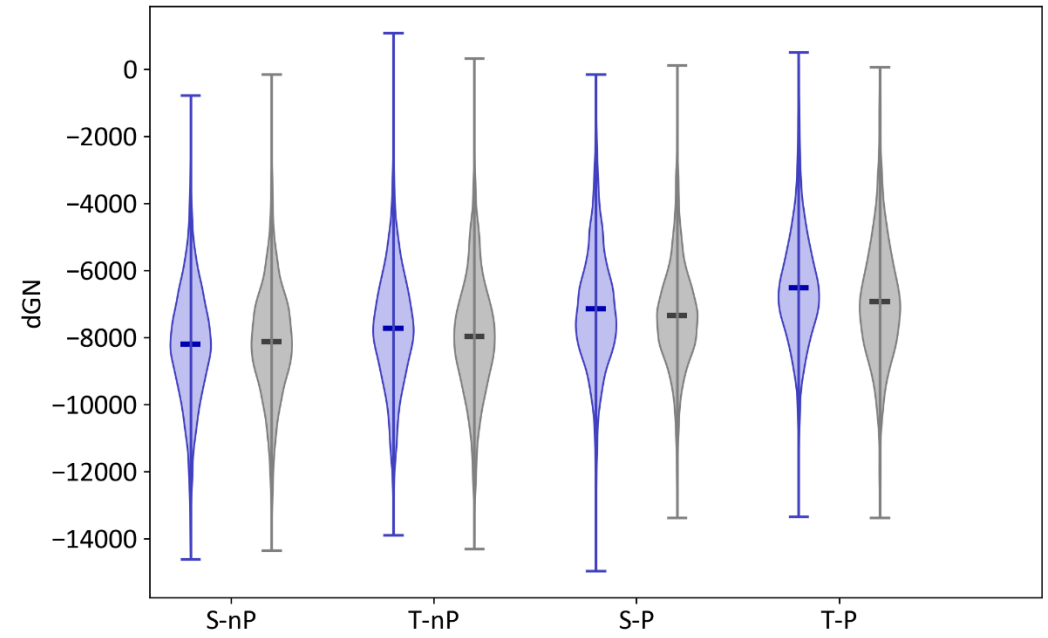

Figure 39. Native state free energy change predicted for phosphorylation sites with phosphomimetic substitution (S → D, T → E)

	Observed ΔG (kcal/mol)	Simulated ΔG (kcal/mol)	ΔG difference (kcal/mol)	ANOVA p-value
S-nP	-8.20	-8.12	0.08	0.079
T-nP	-7.73	-7.51	0.22	2.62E-04
S-P	-7.14	-7.31	-0.17	2.75E-05
T-P	-6.51	-6.92	-0.41	3.26E-13

Table 8. Thermodynamic descriptors calculated with COREX/eSCAPE for phosphorylation sites with +1 site substitution (P ↔ nP)

	Before-phos ΔG (kcal/mol)	Phosphomimetic ΔG (kcal/mol)	ΔG difference (kcal/mol)	ANOVA p-value
S-nP	-8.20	-8.12	0.08	0.051
T-nP	-7.73	-7.96	-0.23	1.49E-04
S-P	-7.14	-7.34	-0.20	2.15E-07
T-P	-6.51	-6.93	-0.42	1.47E-11

Table 9. Thermodynamic descriptors calculated with COREX/eSCAPE for phosphorylation sites with phosphomimetic substitution (S → D, T → E)

substitution increased native state free energy by 0.08kcal/mol for S-nP class, it also decreased native state free energy by 0.23kcal/mol for T-nP class, 0.20kcal/mol for S-P class, and 0.42kcal/mol for T-P class. This indicates the thermodynamic consequences of phosphorylation is affected by not only the identity of phosphorylated residue itself but also by associated environment, including +1 residue.

-3.3.3. Phosphorylated S/T-P sites have different charge / PII distribution

To obtain more concrete evidence about class-specific consequences of phosphorylation, I focused on two familiar properties identified in section 3.3.1: charge and PII propensity. It is known that the charge content of protein is indicative of the general conformation of given sequence, spanning from globules for lightly charged proteins to swollen coils for heavily charged proteins (155). On the other hand, PII propensity governs more local tendency of adopting PII helix conformation, an extended conformation which is a characteristic of both IDRs and phosphorylation sites (107). Both are fundamental properties for dynamics of protein, and IDRs are known to be biased towards higher ends for both scales.

There was a difference in change content (Figure 40) between phosphorylation site classes. S/T-nP sites and tyrosine phosphorylation sites have relatively higher amount of charged amino acids, making the median value to fall in region 2, a boundary between globular region 1 and random coil region 3. Addition of single phosphate pushes the distribution upward, which makes the median to fall in region 3. In contrast, S/T-P sites start from region 1, more collapsed region, and it moves to region 2 after a single phosphorylation.

I already mentioned that there was a PII value difference between phosphorylation site subclasses (Section 3.3.1), and this difference also affects the PII propensity of simulated phosphorylated substrates. S/T-P sites have higher PII values, up to 0.1 higher than its non-proline counterparts, and moves upward after phosphorylation. Our PII scale predicts threonine phosphorylation would make larger changes in PII propensity, while there are several reports which phosphoserine may also contribute to PII formation (107). On the other hand, other phosphorylation site classes showed substantially lower increase of PII propensity.

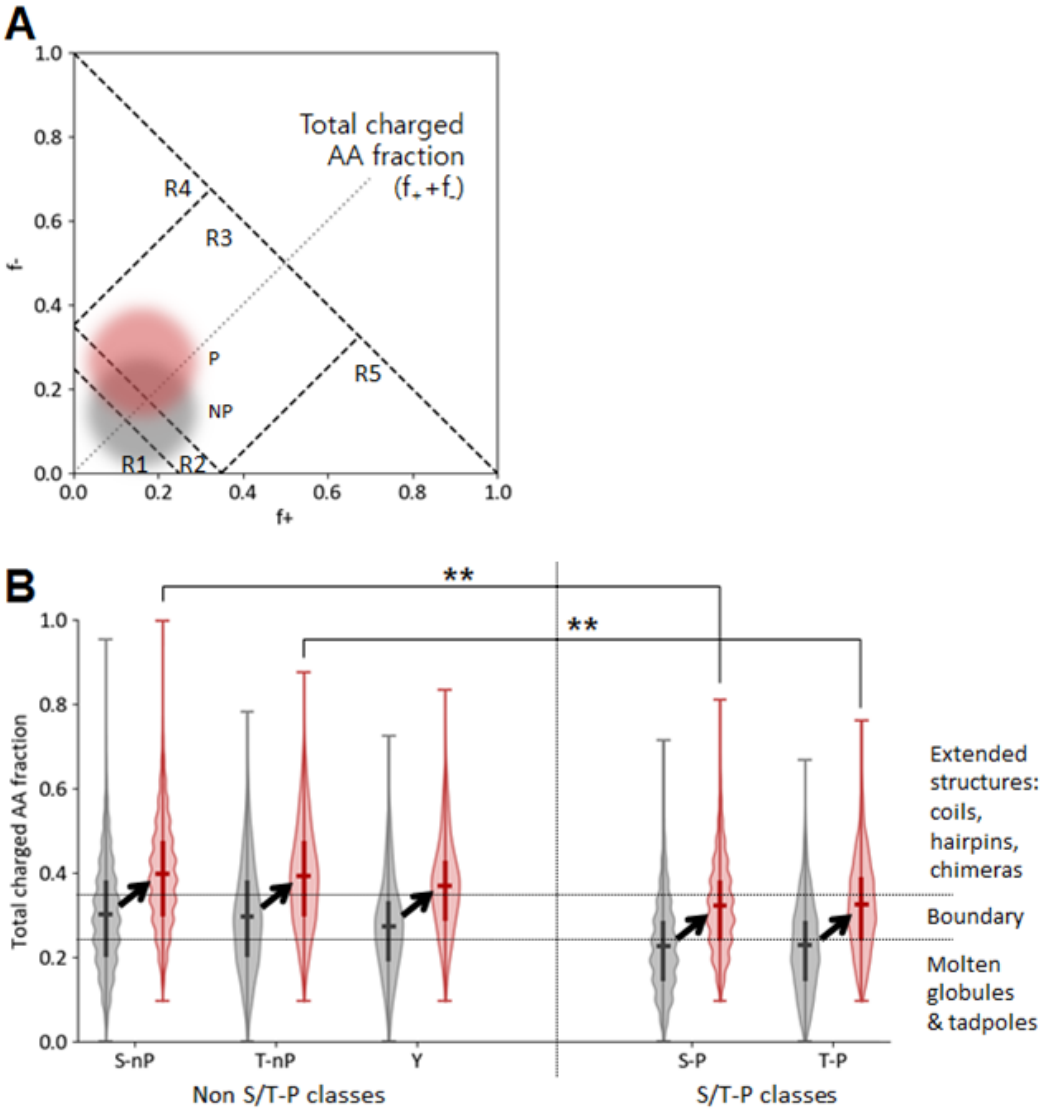

Figure 40. Change of charge content caused by phosphorylation and its effect on thermodynamic environment

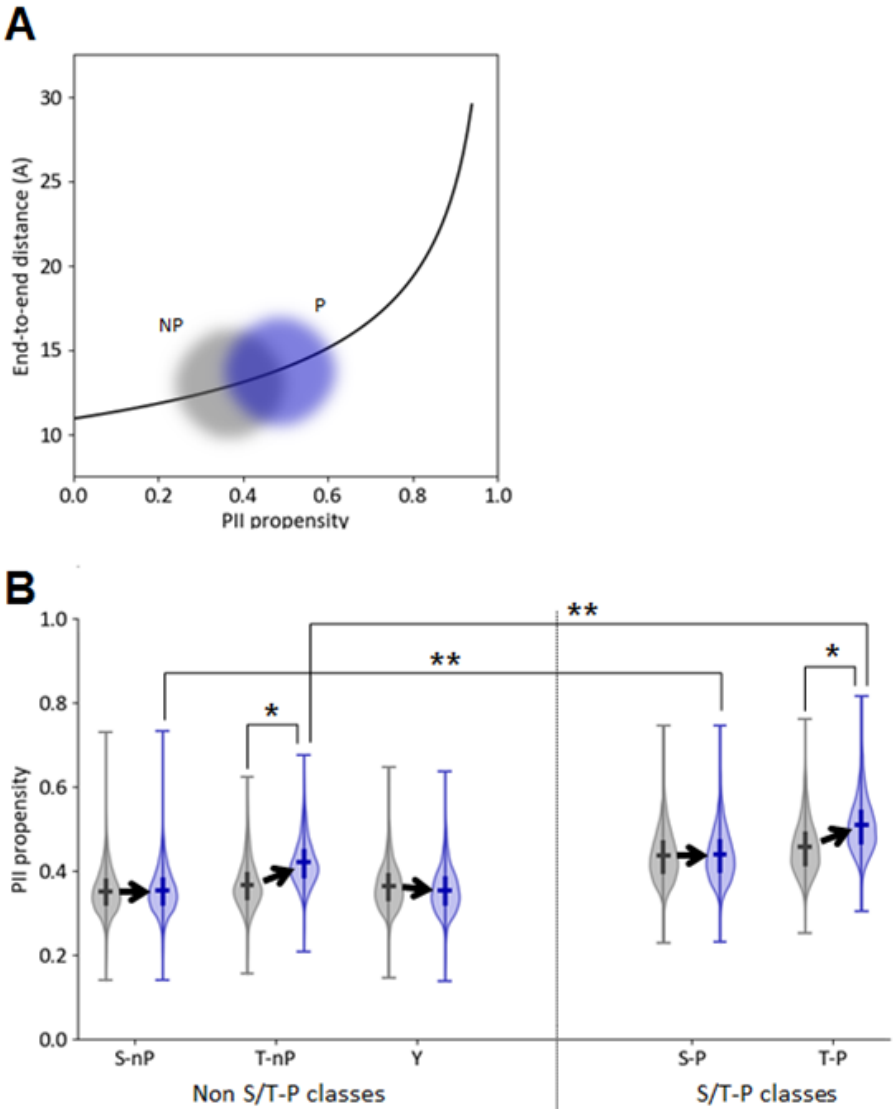

Figure 41. Change of PII propensity caused by phosphorylation and its effect on end-to-end distance

-3.3.4. Phosphorylated S/T-P sites have extended peptide conformation

Results from previous section could be utilized to calculate peptide end-to-end distance (section 3.2.3), a tangible structural property of peptide. To provide a better idea about how the different phosphorylation classes and its phosphorylation affects peptide dynamics differently, I produced a feature space (Figure 42) defined by PII propensity and charge content, and projected the observed charge content / PII propensity changes into this space. This feature space clearly demonstrates differences between S/T-P sites and other phosphorylation classes

All three different phosphorylation classes have different end-to-end distance before and after phosphorylation (Figure 43). Average end-to-end distances of 29AA peptide with a phosphorylation site at the center before phosphorylation were 35.2Å with S/T-P site, 33.7Å with S/T-nP site and 33.8Å with tyrosine phosphorylation site respectively. Simulation of phosphorylation of center residue increased end-to-end distance of given peptide by 0.63Å for S/T-P site, 0.41Å for S/T-P site and 0.24Å for tyrosine phosphorylation sites respectively, thereby increasing the existing difference between S/T-P sites, S/T-nP sites and tyrosine phosphorylation sites.

It was also remarkable that threonine phosphorylation had much stronger effect on end-to-end distances (Figure 44). It is predicted that phosphorylation of threonine residues could trigger end-to-end distance increase by 1.3Å, while phosphorylation of serine or tyrosine residues resulted in minor change (~ 0.2Å). It is probable that the differences between conformational change induced by serine phosphorylation and threonine phosphorylation is exaggerated, as experimentally measured end-to-end distance increases induced by phosphorylation of serine and threonine (105) were not as dramatically different as this result. However, along with phosphomimetic simulation result provided in section 3.3.2, this result provides an implication that there is measurable differences between serine and threonine phosphorylation, which might manifest as differences in functional mechanism and consequences.

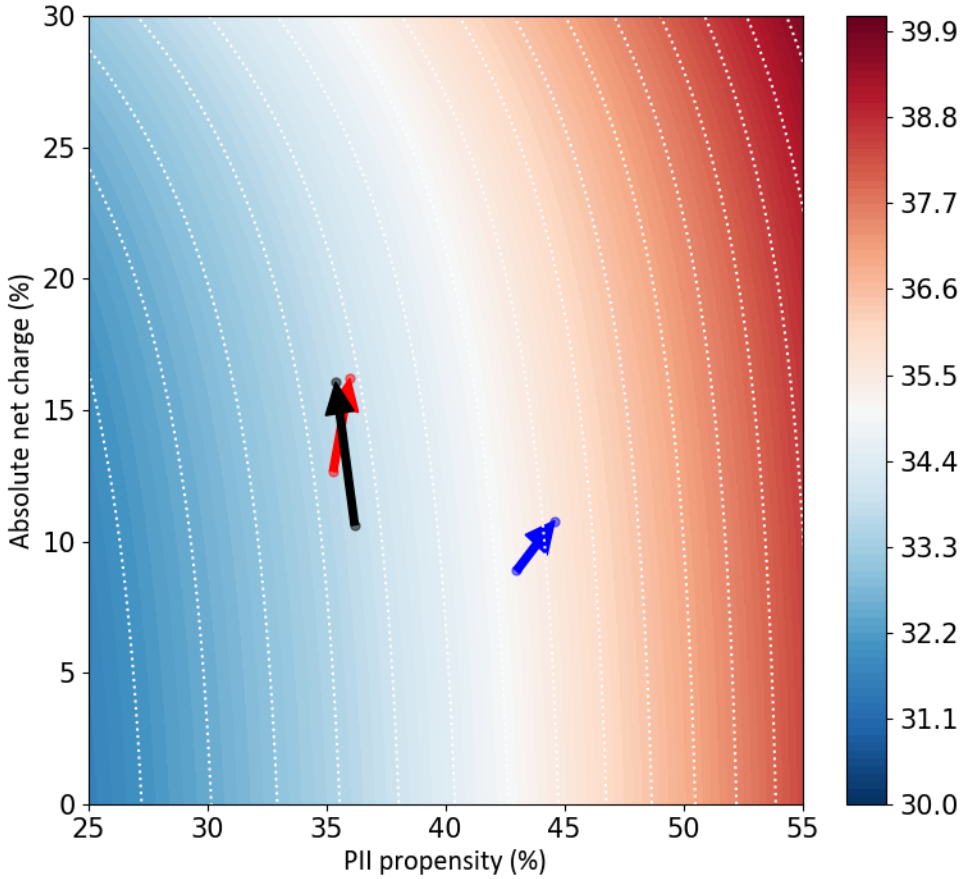

Figure 42. Feature space defined by PII propensity and charge content with average PII / charge values calculated for each phosphorylation classes

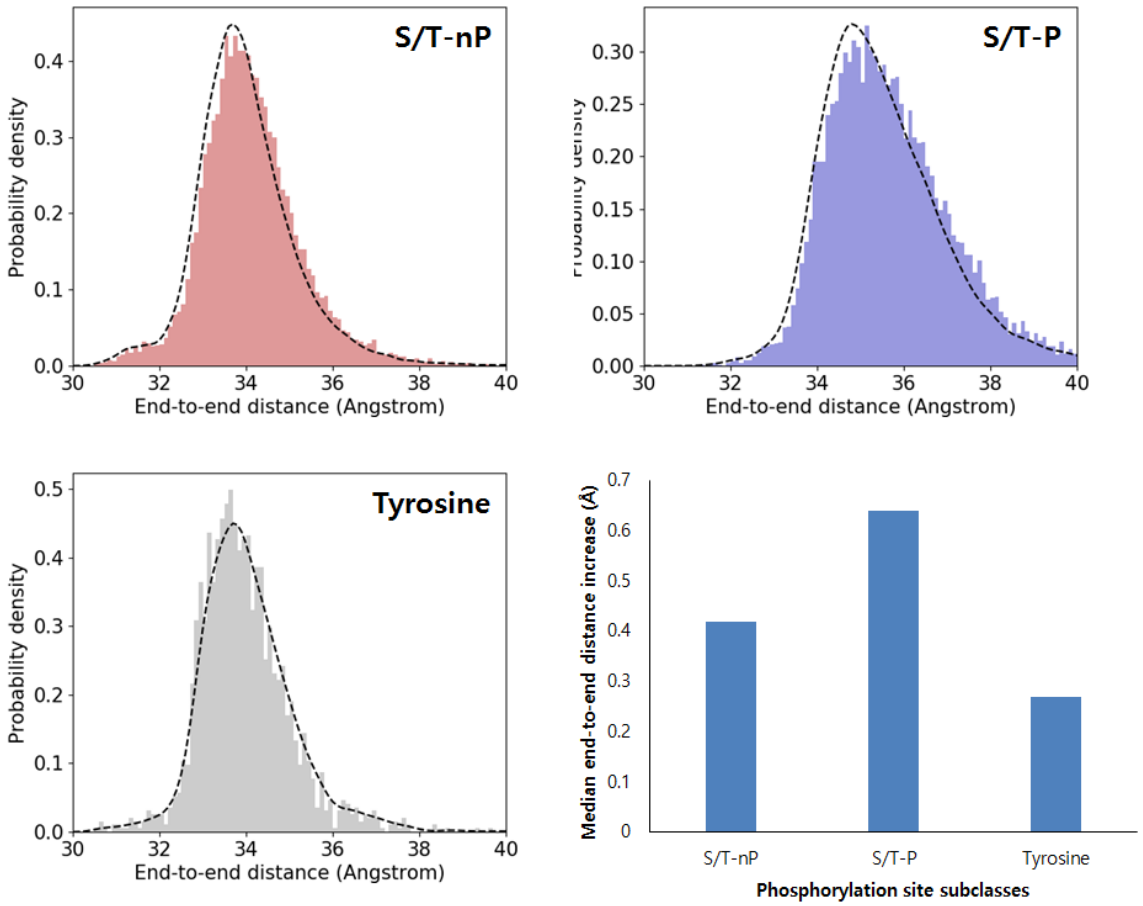

Figure 43. Phosphorylation site subclasses defined with +1 Proline show higher end-to-end distances than other subclasses

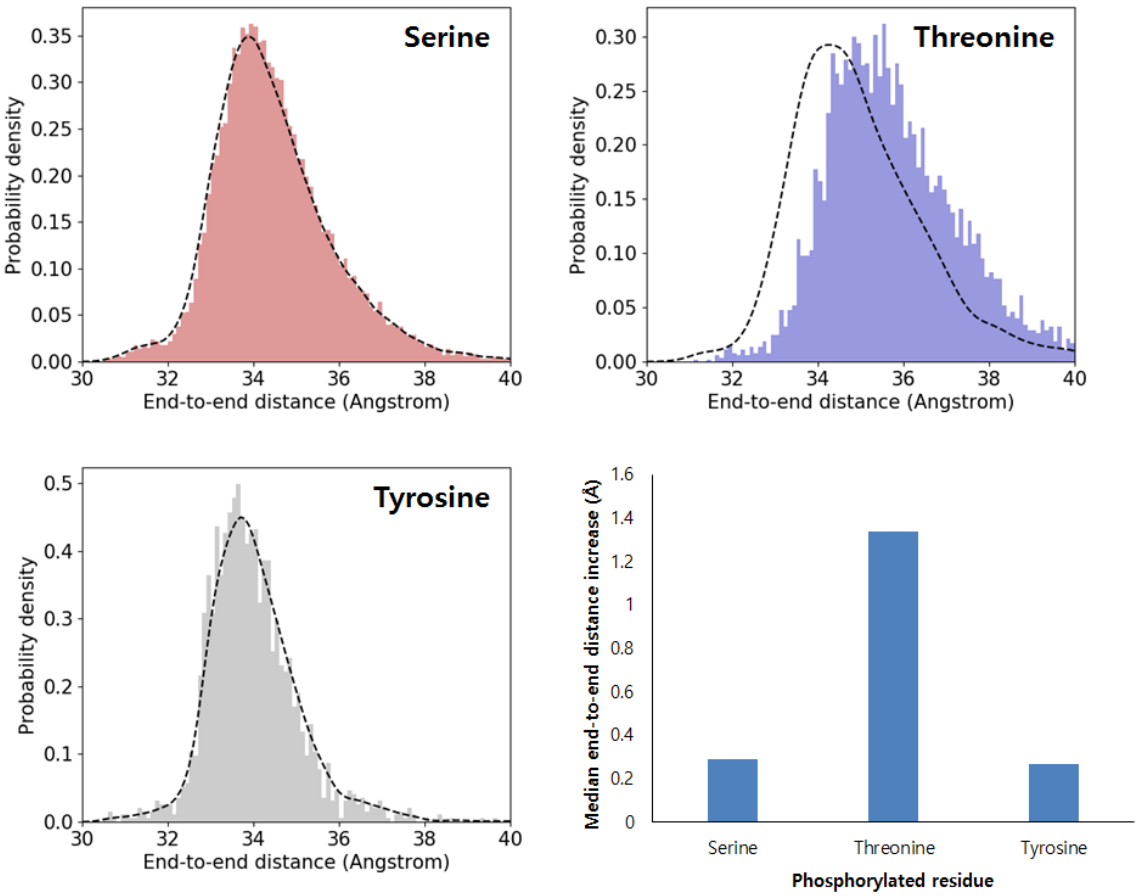

Figure 44. Threonine phosphorylation has a stronger effect on end-to-end distance increase than do serine / tyrosine phosphorylation

-3.3.5. Distribution of S/T-P sites in protein is not random

Phosphorylation sites are likely to have another phosphorylated neighbor nearby. The probabilities of finding phosphorylated neighbor were ~40% within ±5AA region and ~50% within ±10AA region (Figure 45) which is much higher than that predicted from random distribution. This denotes an additional analysis of distribution patterns and its influences on thermodynamic environment is required to avoid my analyses being incomplete.

S/T-P sites and other phosphorylation sites show different preferences towards the class of neighboring phosphorylation site and the distance between neighbors. It was found that phosphorylation sites are likely to have a nearest phosphorylated neighbor of the same class (Figure 46): the probability values were 13% higher for S/T-nP class, 68% higher for S/T-P class, and 8.93-fold higher for tyrosine phosphorylation class than those predicted from null hypothesis respectively. Also, while S/T-nP sites and tyrosine phosphorylation sites preferred nearest phosphorylated neighbor at ±2 sites the most, S/T-P sites preferred it at ±4 sites (Figure 47). Except for +1 site for S/T-P sites (which is fixed to proline so could not be phosphorylated), distribution of phosphorylated neighbors was largely symmetrical – showing no bias towards either N'-terminal or C'-terminal sides.

Again, implementing COREX/eSCAPE and phosphomimetic simulation suggests this different distribution pattern might induce different consequences after phosphorylation (Figure 48). Additional phosphomimetic simulation of nearby phosphorylation sites (phosphorylation site at the center is already substituted with D/E) within ±5AA region further decreased folded state free energy by 0.08kcal/mol and 0.16kcal/mol in average for S-P and T-P sites respectively. On the other hand, same simulation resulted in either increase of free energy by 0.12kcal/mol (S-nP sites) or no significant change of free energy (T-nP sites). Coupled with free energy change caused by single phosphorylation (section 3.3.2), this result deepens the contrast between S/T-nP class and S/T-P class: while multiple phosphorylation around S-nP sites increase the free energy by 0.1~0.2kcal/mol, multiple phosphorylation around T-P sites decrease the free energy by ~ 0.6kcal/mol (both compared to non-phosphorylated substrates).

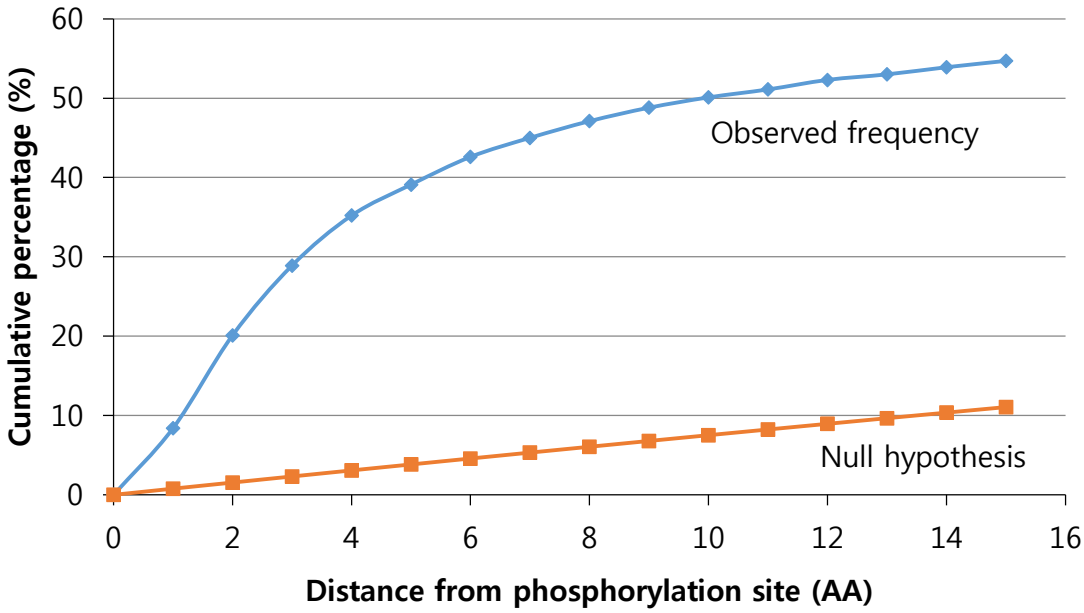

Figure 45. Probability of finding another phosphorylated neighbor within given distance from phosphorylation site

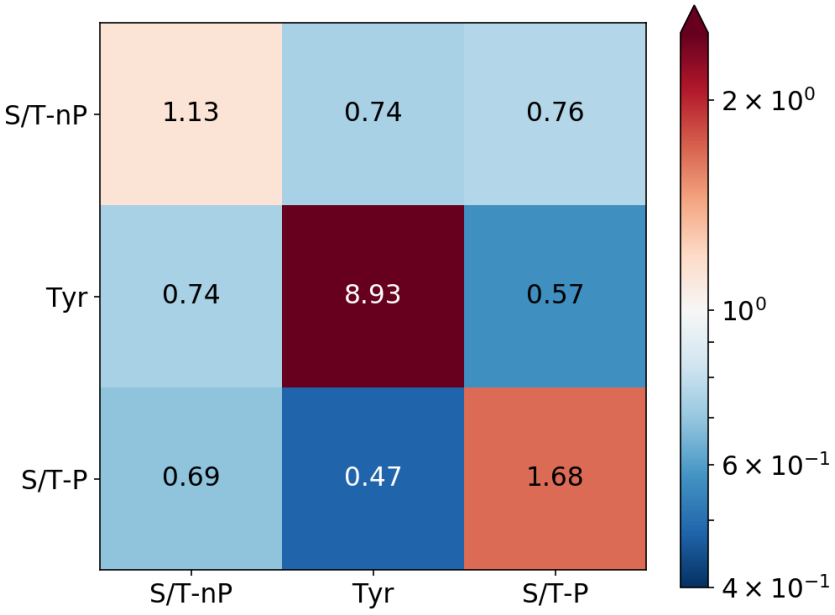

Figure 46. Class-specific statistics of neighboring phosphorylation site pairs

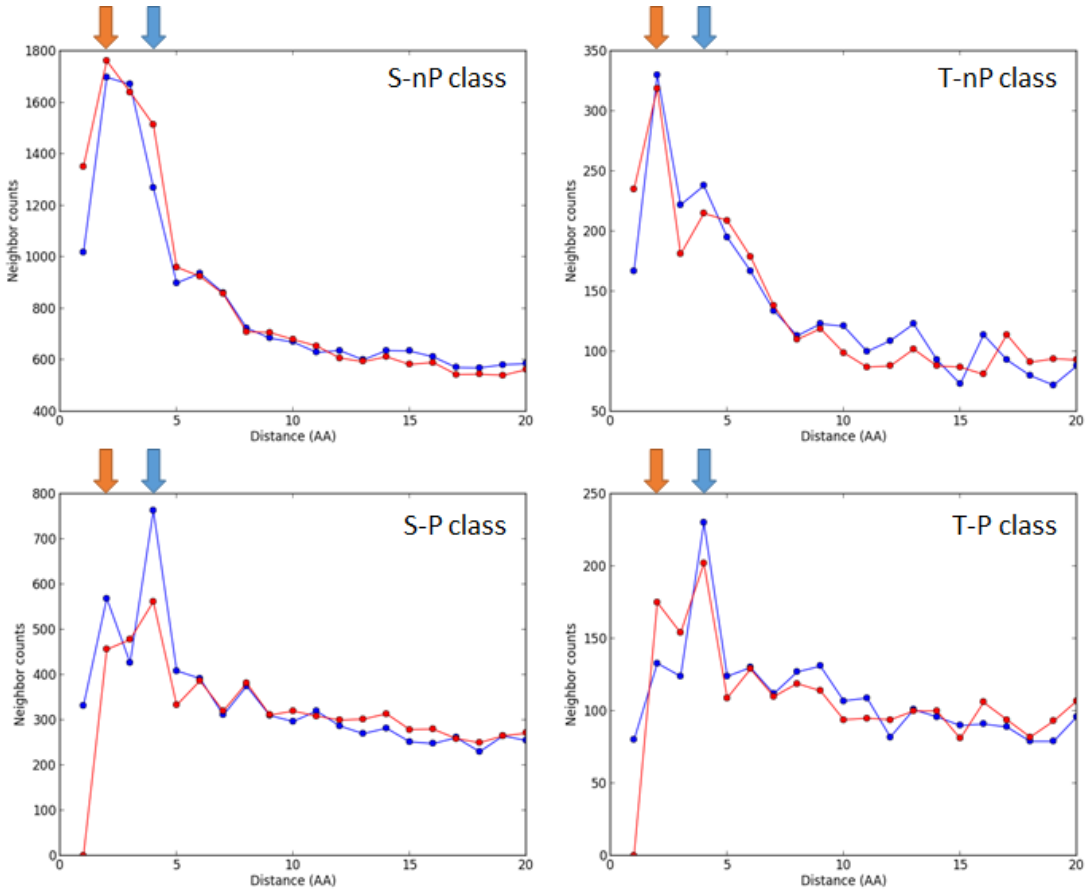

Figure 47. Distribution of distance between nearest phosphorylated neighbors (Light blue arrow: ±4AA site. Orange arrow: ±2AA site. Blue dot & line: distribution of phosphorylated neighbors on N'-terminal side. Red dot & line: distribution of phosphorylated neighbors in C'-terminal side)

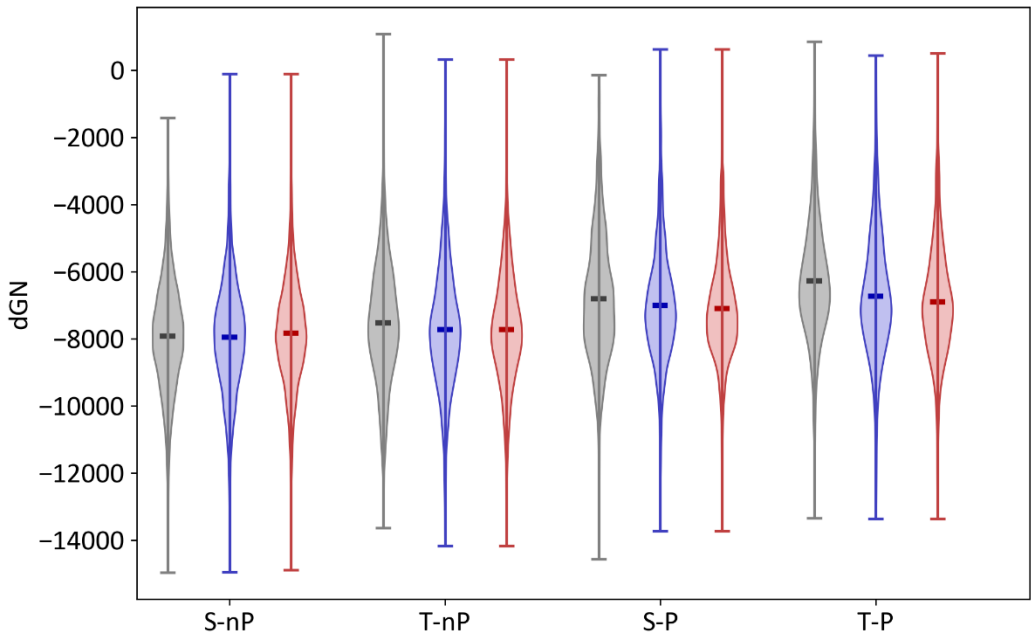

Figure 48. Native state free energy change predicted for phosphorylation sites with singular / multiple phosphomimetic substitution (S → D, T→ E) (Gray: non-phosphorylated state. Blue: single phosphorylated state (at the center). Red: multiple phosphorylated state)

-3.3.6. S/T-P sites are strongly associated with IDRs

Another feature of S/T-P sites should be assessed is its relationship with IDR. While protein phosphorylation itself is clearly associated with IDR, S/T-P sites show even stronger association with intrinsic disorder and flexibility. Almost 80% of all S/T-P sites are found on predicted IDRs, compared to 50% for other phosphorylation sites (Figure 49). Serine and proline amino acids are all known to be disorder-promoting (156), but when compared to the distribution of non-phosphorylated SP / TP dipeptides, accumulation of S/T-P sites were indeed significant (Figure 49).

One thing should be noted is that the fraction of S/T-P sites within phosphorylation sites shows positive correlation with biological complexity (Figure 50B), just as fraction of IDR did (157). Fractions of S/T-P sites in bacteria are in the range 1.5 ~ 10% (Figure 50A), which is much lower than that found in human. It also does not deviate significantly from the frequency of proline (~5%) in the bacterial proteome. However, S/T-P sites become more frequent in eukaryotes, and it showed roughly a linear correlation with the log of cell types. Interestingly, within eukaryotic level, while the ratio of disordered region itself is virtually not correlated with fraction of S/T-P sites (Figure 50C), frequencies of disordered binding region show meaningful correlation (Figure 50D).

[3-4] Discussion

All these results indicate S/T-P sites are associated with different biophysical properties, both local and domain-wise, and this difference may result in different biophysical consequences.

I successfully identified the 'hidden' horizontal information embedded in the sequence which provide information which distinguishes S/T-P sites from non-phosphorylated SP / TP dipeptides. This was consistent with our hypothesis that the positive markers of S/T-P sites are embedded as biophysical properties but not as a discernible sequence feature, which could be which could be conserved substantially strongly than the vertical information in IDRs (Figure 51): more diffuse nature of horizontal properties allow those not to be easily changed by missense mutations frequent in IDRs and low-complexity regions.

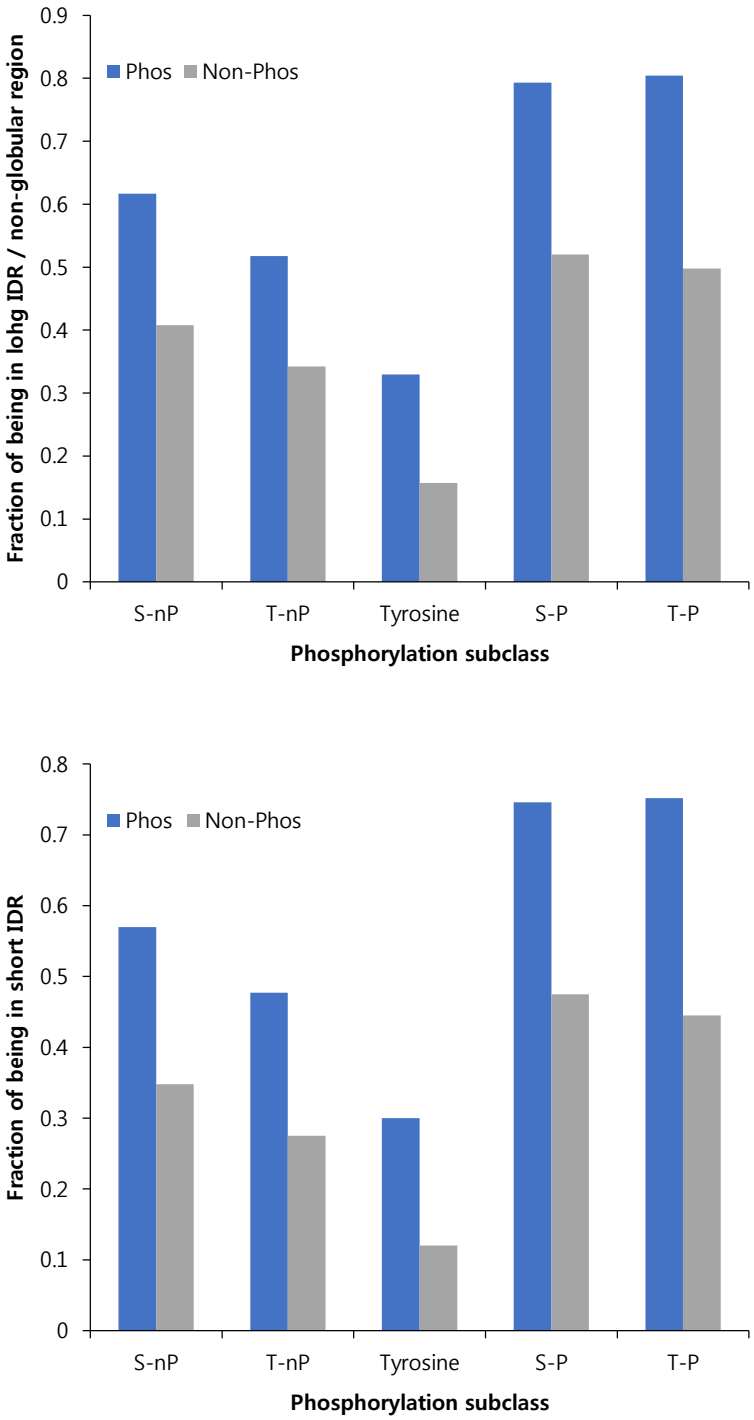

Figure 49. Fractions of phosphorylation sites & non-phosphorylated S / T / Y located in IDRs (Upper panel: long IDRs & globular regions. Lower panel: short IDRs & linkers)

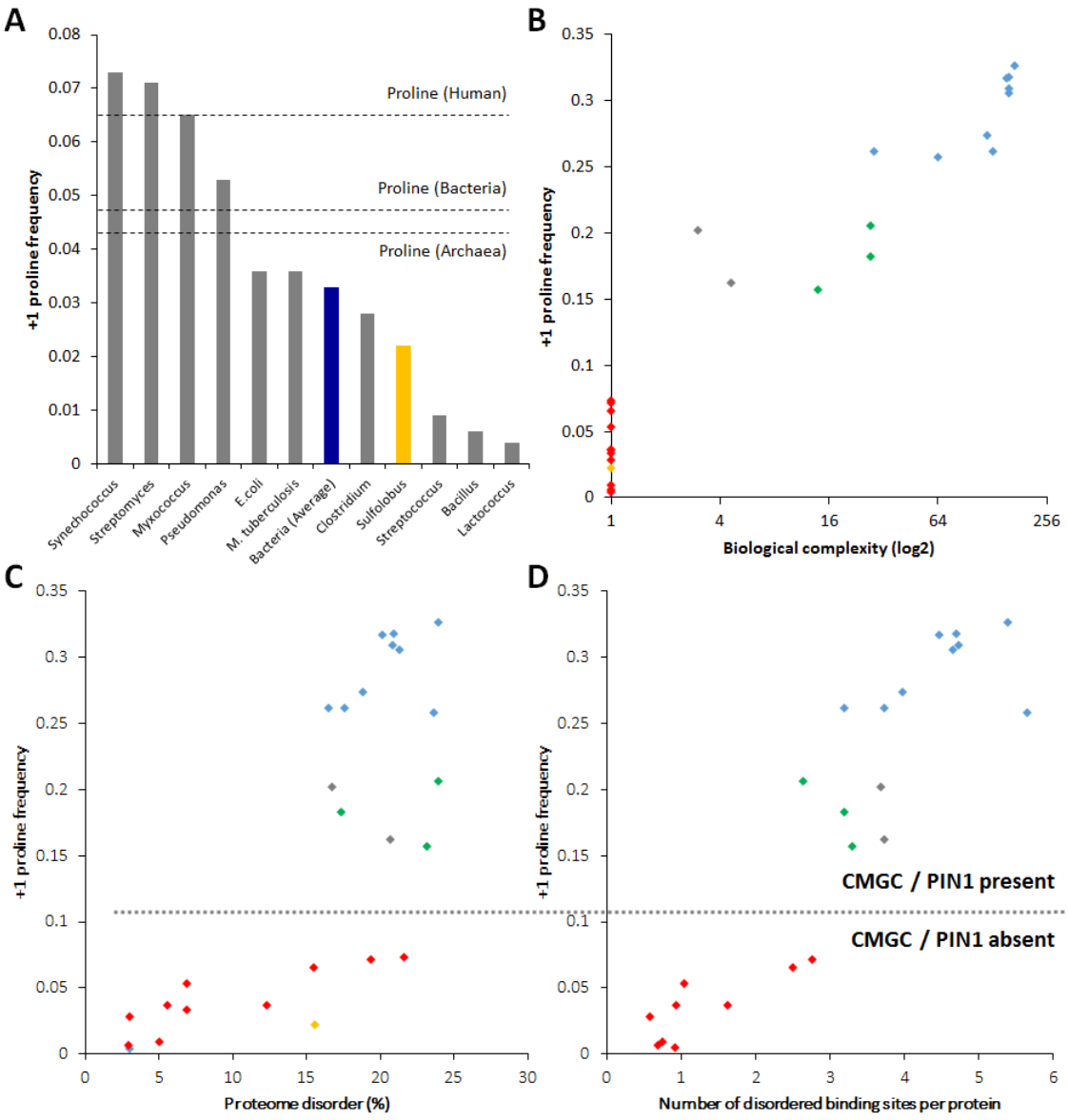

Figure 50. Correlation of S/T-P ratio with biological complexity / IDR-related properties (A: +1 proline frequency found in prokaryotic phosphorylation sites. B: correlation between biological complexity (number of cell types) and +1 proline frequency. C: correlation between proteome disorder content and +1 proline frequency. D: correlation between average number of disordered binding sites per protein and +1 proline frequency)

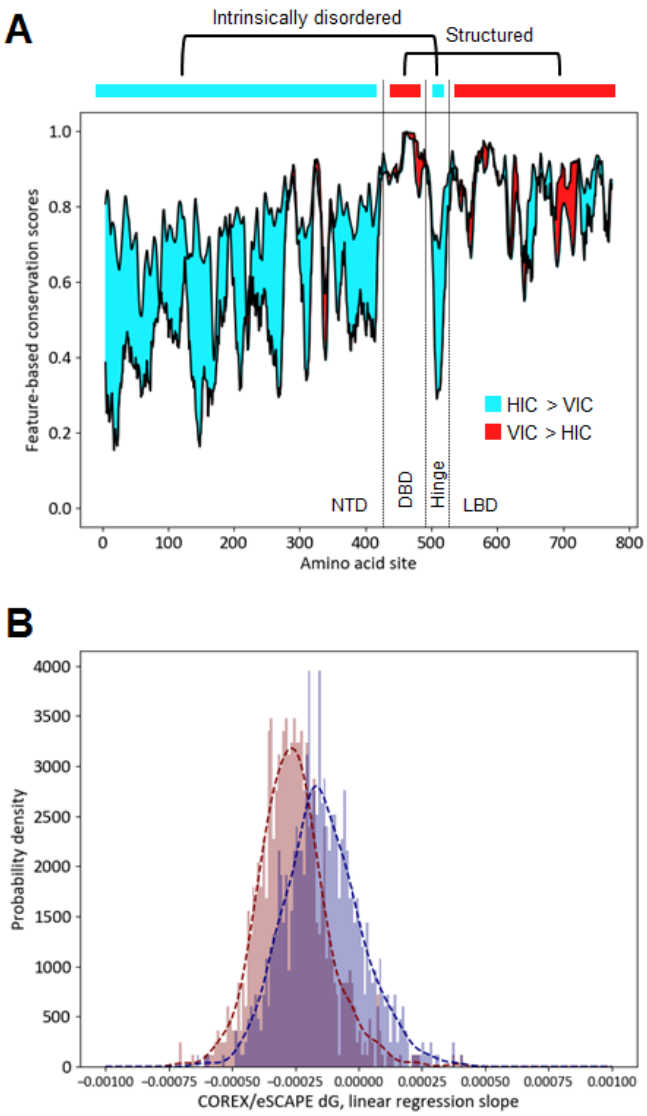

Figure 51. Conservation of native state free energy (horizontal information) and amino acid sequence (vertical information) in IDR and folded regions of human glucocorticoid receptor (GR / NR3C1) (A: Site-specific difference between degrees of conservation of horizontal & vertical information calculated for human glucocorticoid receptor (GR) and its orthologs. B: Correlation coefficients between free energy and conservation score is calculated for ortholog alignments of 835 different transcription factors (see Figure 53)

Also, it was indicative of the involvement of conformational equilibrium of substrates in phosphorylation - but it was not clear at this point whether this substrate thermodynamics matters more in configuration of binding competent structures ('before phosphorylation') or in adaptation of different conformations upon phosphorylation ('after phosphorylation') or possibly both.

Regarding this question, it should be noted again that biophysical properties which differs between phosphorylation sites and non-phosphorylated sequence are not completely the same between phosphorylation site classes, which are indicative of possible mechanistic difference between phosphorylation site classes. For example, S/T-P sites are enriched of PII helix-promoting signals and C'-terminal alpha helix promoting signals (Figures 30, 31). Also, S/T-P sites have higher native state free energy and it is predicted to be decreased upon phosphorylation. Along with previously reported behaviors of serine / threonine phosphorylation of promoting aforementioned secondary structures (62, 107), and behaviors of proline (section 1-4), these signals suggest there is a basal mechanism of inducing specific conformations upon phosphorylation shared between S/T-P sites. On the other hand, other phosphorylation sites show strong biases towards properties associated with exposure or protein-protein interaction, including hydrophobicity, polarity, accessible surface area and charges (Figures 22-23, 36-37), denoting that the embedded information is poised to kinase-substrate binding. This implies contribution of thermodynamics in phosphorylation would be substantially different depending on phosphorylation site class.

It was possible to demonstrate that the environments around different phosphorylation classes would result in different consequences after phosphorylation. Simulating phosphorylated status by the substitution of S/T residues to D/E residues suggested the thermodynamic effects of S/T-P site phosphorylation, including single phosphorylation and multiple phosphorylation, could be substantially different or even opposite from those of S/T-nP site phosphorylation (Figures 39, 48). As biophysical characteristics of pS/pT is not exactly the same as 'proxy' D/E residues, the exact degrees of thermodynamic changes caused by phosphorylation could not be precisely modeled: still, this simulation result was clear enough to show the introduction of same negative charge could produce more than one effects.

Also, different charge and PII propensity of phosphorylation classes would result in different end-to-end

distance increase. It is notable that this extension is highly dependent on the properties of non-phosphorylated state: suggesting the existence of +1 proline, which substantially increases the PII propensity of local peptide, would be the determinant of end-to-end distance increase induced by phosphorylation. On the other hand, higher charged amino acid frequency around other types of phosphorylation sites allows local net charge of phosphorylated substrates to be significantly higher, but the value was not in the range which induces substantial elongation of peptide, which ultimately leads to the smaller end-to-end distance increase. The implication of these results is that S/T-P sites would function differently from other phosphorylation sites, by exploiting the combination of inherent properties of phosphorylation, proline and nearby peptide environment.

Different distribution patterns of S/T-P phosphorylation sites are also indicative of mechanistic differences between phosphorylation site classes. S/T-P phosphorylation sites tend to keep distance from its nearest phosphorylated neighbor, and prefer another S/T-P site over other phosphorylation sites as a nearest neighbor (Figures 45-48) - which might be responsible for further decrease of folded state free energy by multiple phosphorylation (simulated by D/E substitution) predicted only for S/T-P sites. Although speculative, this data might indirectly support the hypothesized role of S/T-P sites as a conformational switch. First, identified distribution pattern of S/T-P sites enlarges the peptide region which could be directly affected by phosphorylation. Ignoring allosteric effects which are mostly protein-specific, general conformational effects of protein phosphorylation are largely limited to the vicinity of phosphorylation site (158). This suggests sparsely distributed S/T-P sites could simply affect not only large number of IDRs but also wide range of peptides which are intrinsically disordered. On the other hand, effects of phosphorylation could interfere with each other: for example, phosphorylation at +2/+3 sites relative to the N'-terminal of alpha helix is known to destabilize the helix (159), clashes with the proposed alpha-helix inducing function of S/T-P site phosphorylation. While this could potentially function as a two-step regulation mechanism, the possibility of simply interfering with the conformational effects could not be excluded at this point.

Proposed conformational mechanism and predicted thermodynamic shift associated with S/T-P site phosphorylation so far imply the observed positive correlation between S/T-P site frequency and IDR content

of eukaryotic proteome is not coincidental. Instead, it is possible that S/T-P site phosphorylation co-evolved with proteome as an adaptation mechanism which allows reversible modulation of IDRs in a coarse-grained manner. This hypothesis is supported by not only the strong association of S/T-P sites with various IDRs but also the fact that both enzymatic machineries associated with S/T-P sites, CMGC kinase family and PIN1 isomerase, are of eukaryotic origin (126, 160). At this point, general lack of information regarding phosphorylation sites prevents from proper evaluation of this hypothesis, but further mechanistic analysis of individual S/T-P sites and investigation of evolutionary trails of phosphoproteome would provide an insight about this implication.

Yet, phosphorylation would be just one of the specific case of more general problem of identifying determinants of biophysical / biochemical interactions targeted to IDRs. The methodology I used here to analyze phosphorylation sites could be also implemented to assess the properties of other PTM sites, intermolecular interactions or other processes which could possibly affect the conformational ensemble. This re-formulation of protein sequences with 'thermodynamic proxies' might identify the embedded information implicated with specific biophysical mechanism and consequently give us a better understanding about IDR regulation in general.

[4] PHOSforUS: a biophysical property-based phosphorylation site predictor

[4-1] Introduction

In this chapter, I'd like to demonstrate how the differences identified from the previous chapters could be applied to design an effective phosphorylation site predictor. The content in this chapter is largely based on my paper published in PNAS (148) (under revision).

Despite its perceived importance, only a small number of phosphorylation sites are experimentally validated and studied, which produces a substantial knowledge gap which prevents understanding of how phosphorylation mediates biological processes. This gap is exacerbated by the fact that the majority of phosphorylation sites are found in IDRs: due to its divergent sequence, it is increasingly difficult to identify possible phosphorylation sites based on sequence comparison with already identified sequences. Moreover, phosphorylation is both transient and reversible, which make reliable identification of phosphorylation sites more complicated. This is reflected in the low degree of consensus (161) between several major annotation databases, including SWISS-PROT, Phospho.ELM and PhosphoSitePlus (25 35, 117).

Heuristics augmenting limited amount of experimentally validated annotations, such as sequence motifs and position-specific weight matrices (PSWM), has been developed to address aforementioned knowledge gap. Coupled with more sophisticated machine learning techniques such as artificial neural network (ANN) or support vector machine (SVM), a moderate success in predicting novel phosphorylation sites was achieved: however, due to the substantial variability in the deducted consensus patterns, development of prediction tools based on this approach has been slowed down significantly (162)

Based on the findings discussed in previous chapters, I hypothesized that the conformational equilibrium of protein would be the determinant of not only the function of phosphorylation sites but also the kinase specificity. To test this hypothesis, we developed a framework which utilizes both site-specific sequence elements conserved at particular position (vertical information) and ensemble-averaged properties which are conserved along the small window of sequence (horizontal information), which resulted in a predictor, PHOSforUS, which outperforms most of existing phosphorylation site predictors. The results show that the

consideration of horizontal properties which encodes equilibrium fluctuations contributed to the increased predictive performance. PHOSforUS is currently freely available at https://github.com/bxlab/PHOSforUS.

[4-2] Approaches

-4.2.1. Reference dataset

Canonical human protein sequences were obtained from SWISS-PROT (2018 December Release) (25), a manually curated subset of the UniProt database. Phosphorylation annotations were obtained from SWISS-PROT and PhosphoSitePlus (2018 December Release) (35). True positive sets were generated from SWISS-PROT annotations and low-throughput (LTP) category of PhosphoSitePlus. Sequence fragments with length = 29 (14 residues N-terminal and C-terminal relative to a central phosphorylation site) were collected from these sets and subsequently divided into five subclasses (S-P, S-NP, T-P, T-NP, Y) based on the identity of the center residue and the presence of +1 Pro as its C-terminal neighbor. For example, S-P denotes Ser as the phosphorylatable central residue with presence of the +1 Pro, while S-NP denotes any of the remaining 19 residues at the +1 position. To reduce any possible redundancy, a 50% maximum pairwise sequence similarity filter was applied to these subsets. True negative subsets were generated in a similar way and sequences that shared more than 50% similarity to any phosphorylated sequence (both LTP and HTP) were removed to filter out false positives. Resulting statistics of these sets are shown in Table 10.

For the comparative analysis, we constructed an alternative positive sets which contain no sequence within training set of our predictor, and presumably minimal number of sequences in the training sets of available phosphorylation predictors. From PhosphoSitePlus high-throughput (HTP) subset, we removed sequences that show 50% similarity to any of sequences within SWISS-PROT, Phospho.ELM (117) and PhosphoSitePlus LTP datasets. From resulting positive set (Table 10) and true negative set, we sampled 5 testing sets with 100 positive sites and 100 negative sites randomly to calculate predictor performances and subsequently make a comparison.

Class	Total P-sites			Total N-sites	
	Pre-screening	After screening	Comparative analysis	Pre-screening	After screening
S-P	10348	3024	11842	30170	7373
S-nP	21936	4426	55628	455303	88905
T-P	2688	1176	5028	20943	1762
T-nP	3045	1385	20299	288492	27627
Y	2058	1145	14415	145170	24271

Table 10. Training / testing set statistics utilized for PHOSforUS training & testing

-4.2.2. Feature selection

Features utilized for PHOSforUS were selected from 546 amino acid scales based on information content. These scales were collected from from the AAindex database (152) with the addition of the DisProt scale (153) and experimental polyproline II (PII) (107) propensity scale. To assess information content of a scale, analysis datasets were built with randomly selected 1000 true positives and 1000 true negatives, both coming from the same phosphorylation subset of 29AA length fragments. For every fragment, a weighted average of values from the scale with window size = 9 was calculated for the region of 21 residues centered on the serine / threonine / tyrosine residues, and the information value of each scale was estimated using a naïve Bayes classifier (section 4.2.4) with x10 cross validation. This procedure was repeated x10 with randomly selected datasets and individually tested for each subclasses. For each subclass, an amino acid scale which resulted greater than 0.6 prediction accuracy was retained, otherwise the scale was rejected. In this way, 114 scales were retained from the original set. This number was further reduced to 35 by keeping only one of scale pairs exhibiting an absolute Pearson correlation coefficient greater than 0.8. Finally, manual curation to remove redundant scales based on AAindex descriptors resulted in ten features used in the predictor (Table 11).

Additional features for the predictor were obtained from our unique sequence-based energy prediction tool, COREX/eSCAPE (146). From 28 thermodynamic features (including stability, enthalpy, and entropy of both native and denatured states) eSCAPE calculates, we empirically selected four native state features and four denatured state features whose values and differences seemed to be effective in phosphorylation site prediction. The eight features and differences are listed in Table 12. Thus, a total of 18 features were used in the final predictor.

-4.2.3. Machine learning algorithms

To construct a phosphorylation site predictors based on biophysical properties calculated from the sequence, I utilized two learning methods: naïve Bayes classifier (163) and gradient boosting classifier (164).

Feature ID	Description	Feature type	Reference

GUYH850101	Partition energy	Hydrophobicity / Horizontal	Guy (1985)
MIYS990104	Optimized relative partition energy - method C	Hydrophobicity / Horizontal	Miyazawa-Jernigan (1994)
PRAM900102	Relative frequenct in alpha-helix	Conformation / Horizontal	Prabhakaran (1990)
PALJ810112	Normalized frequency of beta-sheet	Conformation / Horizontal	Palau et al. (1981)
ROBB760105	Information measure for extended	Conformation / Horizontal	Robson-Suzuki (1976)
PPIIPRO	Polyproline II propensity	Conformation / Horizontal	Elam et al. (2013)
ZIMJ680104	Isoelectric points	Vertical	Zimmerman et al. (1968)
FASG760101	Molecular weight	Vertical	Fasman (1976)
GRAR740103	Residue volume	Vertical	Grantham (1974)
RADA880106	Accessible surface area	Vertical	Radzicka-Wolfenden (1988)

Table 11. List of biophysical indices incorporated in PHOSforUS predictor

***eSCAPE* parameter**	**Description**
ΔG_N	Gibbs free energy of folded state
$\Delta H_{ap,N}$	Apolar enthalpy of folded state
$\Delta H_{pol,N}$	Polar enthalpy of folded state
$T\Delta S_{conf,N}$	Conformational entropy of folded state
$\Delta\Delta G$ ($\Delta G_N - \Delta G_D$)	ΔG difference between folded & unfolded state
$\Delta\Delta H_{ap}$ ($\Delta H_{ap,N} - \Delta H_{ap,D}$)	ΔH_{ap} difference between folded & unfolded state
$\Delta\Delta H_{pol}$ ($\Delta H_{pol,N} - \Delta H_{pol,D}$)	ΔH_{pol} difference between folded & unfolded state
$\Delta T\Delta S_{conf}$ ($T\Delta S_{conf,N} - T\Delta S_{conf,D}$)	$T\Delta S_{conf}$ difference between folded & unfolded state

Table 12. List of eSCAPE thermodynamic parameters incorporated in PHOSforUS predictor

Naïve Bayes classifier is a group of probabilistic classifier based on Bayes' theorem, which allows to calculate posterior probability from likelihood and priors. The probability of a sample with feature vector $x = \{x_1, \dots, x_n\}$ belonging to class C_m could be calculated as:

$$p(C_m|x) = \frac{p(x|C_m)p(C_m)}{p(x)} \quad \dots (20)$$

Naïve Bayes classifier makes two assumptions: each feature is independent of each other, and each feature has an equal weight - hence ' Naïve '. Therefore, by using chain rule, the expression could be re-written as following equation 21:

$$p(C_m|x) = \frac{1}{p(x)} p(C_m) \prod_{i=1}^{n} p(x_i|C_m) \quad \dots (21)$$

The scaling variable 1/p(x) is constant if values of individual features are known. Naïve Bayes classifier constructs hypotheses for each classes and selects a hypothesis which is the most probable (maximum a posteriori rule).

$$\hat{y} = \underset{m \in \{1,\dots M\}}{\operatorname{argmax}} p(C_m) \prod_{i=1}^{n} p(x_i|C_m) \quad \dots (22)$$

In this research, I used Gaussian Naïve Bayes, as most of biophysical features I used are continuous. The model assumes that the distribution of feature values follow a Gaussian distribution, therefore the probability of having specific value given class C_m could be calculated from mean and variance of distribution.

$$\mathrm{p}(x_n = v \mid C_m) = \frac{1}{\sqrt{2\pi\sigma_{n,m}^2}} e^{-\frac{(v-\mu_{n,m})^2}{2\sigma_{n,m}^2}} \quad \dots (23)$$

Naïve Bayes classifier is among the simplest supervised learning methods, and therefore it takes only a small amount of time to train the model. However, it often produces predictive performances comparable to those from more sophisticated learning methods such as support vector machine (SVM) or artificial neural network (ANN). It is also relatively robust to overfitting, which allows to train a model with smaller number of training samples. As phosphorylation site annotation is relatively scarce (especially for threonine / tyrosine phosphorylation sites), Naïve Bayes classifier is implemented to avoid possible overfitting issues.

On the other hand, gradient boosting classifier produces a model which is essentially an ensemble of decision trees constructed by boosting process. Training involves the addition of a weak learner to the existing ensemble one-by-one to reduce the loss function, which is iterated until the predictive performance of ensemble model could not be improved by the addition of learners. For choosing the best learner at each step, it applies steepest gradient approach to simplify the problem.

The classifier model $f(x)$ could be expressed as follows:

$$f(x) = f_0(x) - \sum_{j=1}^{m} \alpha \cdot \gamma_j \sum_{i=1}^{n} \nabla_{f_{j-1}} L\left(y_i, f_{j-1}(x_i)\right) \quad \ldots (24)$$

$$\gamma_j = arg \min_{\gamma} \sum_{i=1}^{n} L\left(y_i, f_{j-1}(x_i) - \gamma \nabla_{f_{j-1}} L\left(y_i, f_{j-1}(x_i)\right)\right) \quad \ldots (25)$$

Here, $f_0(x)$ is the initial model, L(y, f(x)) is a loss function, and α is a learning rate. Small learning rate is known to dramatically improve the generalization ability of given model at a cost of training time (165). Also, properties of individual learners such as depth of tree or total number of nodes are known to affect predictive performances of ensemble model.

Gradient boosting is known for its predictive performances and relatively low computational burden required for training models. Compared to random forest, another ensemble model based on decision trees, gradient boosting is relatively prone to overfitting issues – which could be avoided by proper setting of learning parameters, takes more time to train a classifier model, but could result in better performances and faster prediction using trained models – which makes it suitable for batch prediction of phosphorylation sites.

-4.2.4. Evaluation of predictive performances

10-fold cross-validation was performed to evaluate the sensitivity, specificity, and accuracy of the prediction models. As the true negative set is much larger than the true positive set, random sampling of the true negative set equalized the numbers of true and false positives during the evaluation. Cross-validation was iterated ten times with different true negative sets to minimize sampling error.

The following evaluation metrics were used: Sensitivity (true positive rate), Specificity (true negative rate), Positive predictive value, Accuracy, F1 score and Matthews correlation coefficient (MCC). In the following equations 26 to 31, TP stands for true positive, FN for false negative, FP for false positive, TN for true negative.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad \ldots (26)$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad \ldots (27)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad \ldots (28)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \quad \ldots (29)$$

$$\text{F1 score} = \frac{2TP}{2TP+FP+FN} \quad \ldots (30)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad \ldots (31)$$

[4-3] Results

-4.3.1. Design concept of PHOSforUS

PHOSforUS incorporates both horizontal and vertical information to predict phosphorylation sites, which correspond to the two steps of kinase-substrate recognition equilibrium:

$$S_{BI} \underset{K_{conf}}{\Longleftrightarrow} S_{BC} + E \underset{K_{int}}{\Longleftrightarrow} ES_{BC} \quad \ldots (32)$$

In equation 32, E is the kinase, S is the unphosphorylated substrate, and the subscripts denote the conformations of the substrates – BI for binding incompetent, BC for binding competent. It is important that the binding competent and binding incompetent states are agnostic as to the degree of structure present, only that the energetic barrier exists between the sub-ensemble that could bind to be phosphorylated and the other sub-ensemble that could not. Also, the equation defines two free energy contributions towards protein

phosphorylation: one from the organization of conformational ensemble of substrate (K_{conf}) and one from the binding interaction of kinase and substrate (K_{int}). We assumed that these contributions could be divided and accessed in terms of measurable information values (Figure 52). This scenario suggests both conformational ensemble of the substrate and site-specific sequence elements would encode the information about kinase specificity.

The approach of assessing the contributions to (K_{conf}) encoded in the horizontal information is predicated on previous studies from our group presenting that proteins can be represented as sequences of thermodynamic environments (166) which capture the experimentally observed fluctuations in conformation (both ordered and disordered ensembles) (167). Also, propensities of amino acids in these thermodynamic environments provide sufficient information to match unknown proteins to their environmental profiles (168) which are conserved (169). The importance of these previous findings is that they show that concealed information about the thermodynamics of a chain is still embedded within the sequence itself, which could be accessed by comparing this sequence-averaged ('horizontal' information for other sequences (Figure 52).

Conservation of the horizontal information could be in fact stronger than the conservation of actual protein sequences in some biological contexts. For example, position-wise native state free energy (146) among the members of the intrinsically disordered N-terminal region of the glucocorticoid receptor (GCR / NR3C1) family is significantly conserved, while the sequence of given region is not (150). And this behavior seemed to be a general feature of protein ortholog groups (Figure 51, 53) - suggesting that horizontal information could be conserved to certain degree even in the absence of sequence conservation, which further motivated the combination of both types of information for the prediction of protein properties: prediction of S/T-P sites in particular, as it is associated with minimal site-specific features but multiple biophysical properties.

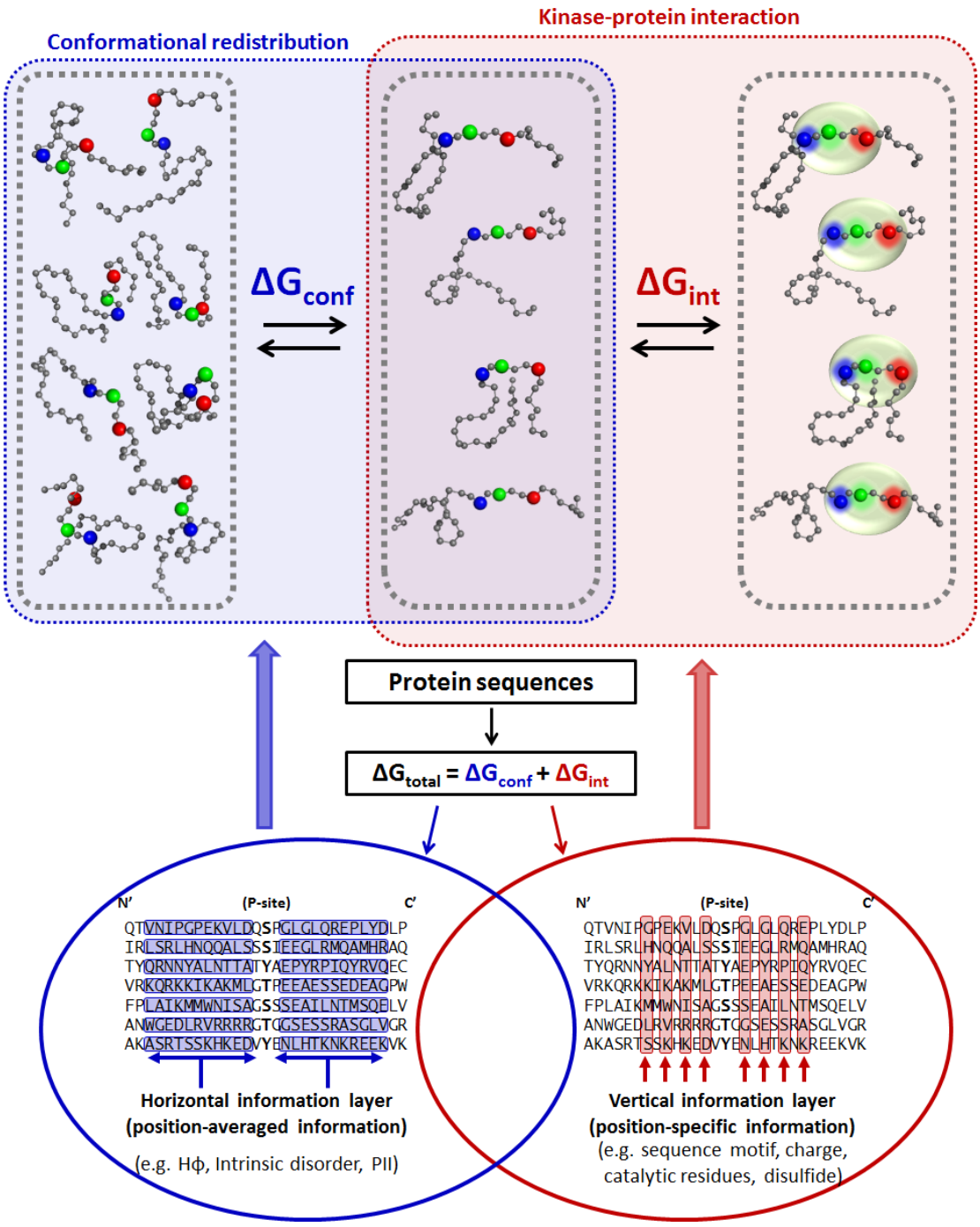

Figure 52. Schematics: horizontal and vertical protein sequence information reflected in the conformational and binding equilibria of kinase-substrate interaction

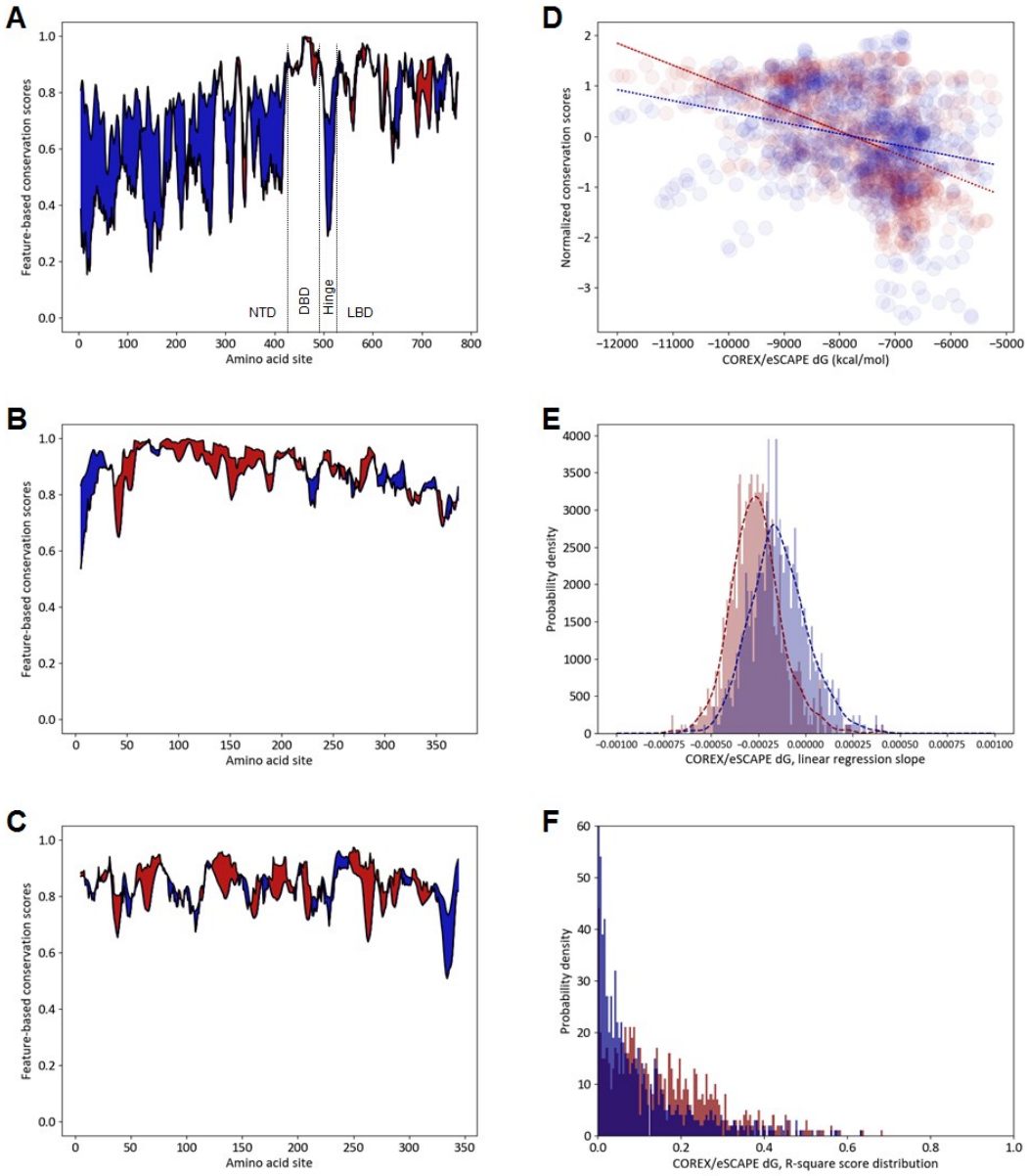

Figure 53. Horizontal information is better conserved than vertical information in IDRs (Red: vertical information, Blue: horizontal information) (A-C: Difference between degrees of conservation of sequence and free energy (ΔG, (5)) calculated for human glucocorticoid receptor (A), actin (B) and rhodopsin (C). D: Correlation between position-specific COREX/eSCAPE ΔG and sequence conservation. E: Distribution of linear regression slopes for 835 different transcription families. F: Distribution of R-square values of linear regression)

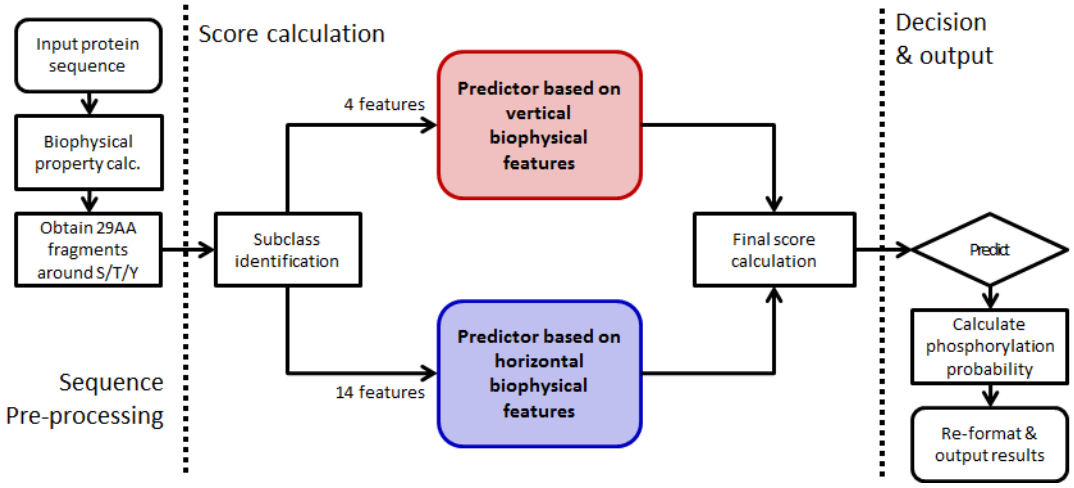

Figure 54. Simplified algorithm architecture of PHOSforUS

-4.3.2. Architecture of PHOSforUS

In the latest PHOSforUS predictor, total of 18 biophysical properties were utilized (Tables 11, 12). These properties were selected to cover the most of biophysical features that are found to be associated with protein phosphorylation while minimizing computational workload. The details of feature selection process are elaborated in section 4.2.2. The property values were calculated for 21 sites (-10 ~ +10 positions) from the input sequence, generating feature vector with 378 property values for each possible phosphorylation sites (S / T / Y residues) found in the input sequence.

PHOSforUS has a nested predictor structure (Figure 54) based on Gaussian Naïve Bayes classifier and gradient boosting classifier (Section 4.2.3). Biophysical property values calculated from the input sequence were first processed with Naïve Bayes subpredictors trained for individual phosphorylation classes. The predictive values generate from subpredictors were applied to downstream metapredictor based on gradient boosting, thereby calculates the probability of being phosphorylated. One important fact here is that the predictor components of PHOSforUS do not take the sequence information directly: while vertical property values are practically a numericized sequence itself, consequent Gaussian Naïve Bayes step takes the input as continuous values, thus the multivariate properties of protein sequences are effectively discarded. This makes PHOSforUS truly a unique predictor which is minimally dependent on exact protein sequence information.

-4.3.3. Predictive performances of PHOSforUS

Predictive performances of each Naïve Bayes sub-predictors (Tables 13-17) were analyzed to measure the contribution of each information values in prediction. Sub-predictors trained with either of two hydrophobicity values (GUYH850101 / MIYS990104) produced the best results for all five classes. Isoelectric point of amino acids (ZIMJ680104), which I used as a proxy of residue charge, resulted in better prediction results for S-nP, T-nP and tyrosine phosphorylation sites, while residue volume (GRAR740103), accessible surface area (RADA880106), alpha-helix propensity (PRAM900102) and PII propensity were

Class S-P	Accuracy	Sensitivity	Specificity	Precision	F1	MCC	AUROC
ZIMJ680104	0.598934	0.549289	0.648578	0.609838	0.577963	0.198858	0.640449
FASG760101	0.644826	0.646998	0.642654	0.644284	0.645597	0.289692	0.698131
GRAR740103	0.680016	0.664929	0.695103	0.685712	0.675114	0.360244	0.743468
RADA880106	0.673697	0.620458	0.726935	0.6945	0.655319	0.34945	0.740499
Vertical features	**0.752725**	**0.738784**	**0.766667**	**0.760029**	**0.749216**	**0.505698**	**0.835519**
GUYH850101	0.752765	0.76722	0.73831	0.745754	0.756271	0.505829	0.834035
MIYS990104	0.759874	0.779226	0.740521	0.750265	0.76444	0.52018	0.840712
PRAM900102	0.619471	0.517615	0.721327	0.649957	0.576162	0.244081	0.663511
PALJ810112	0.664218	0.704265	0.624171	0.652126	0.677157	0.329528	0.724739
ROBB760105	0.702291	0.732148	0.672433	0.690894	0.710886	0.405357	0.774226
PPIIPRO	0.662046	0.530174	0.793918	0.720275	0.610634	0.336092	0.724
ΔG,N	0.683965	0.709795	0.658136	0.674974	0.691914	0.368459	0.748253
ΔH_{ap},N	0.612243	0.539652	0.684834	0.631285	0.581816	0.226914	0.657694
ΔH_{pol},N	0.622749	0.674724	0.570774	0.611197	0.641368	0.246861	0.667824
$T\Delta S_{conf}$,N	0.606635	0.671248	0.542022	0.594464	0.630492	0.215103	0.642729
ΔΔG,N-D	0.693009	0.65158	0.734439	0.710504	0.679719	0.387398	0.762268
$\Delta\Delta H_{ap}$,N-D	0.674171	0.71169	0.636651	0.662028	0.685947	0.349344	0.736206
$\Delta\Delta H_{pol}$,N-D	0.635427	0.590758	0.680095	0.648688	0.618317	0.271969	0.687678
$T\Delta\Delta S_{conf}$,N-D	0.667457	0.627409	0.707504	0.682081	0.653554	0.336035	0.728165
Horizontal features	**0.78207**	**0.792733**	**0.771406**	**0.77631**	**0.784389**	**0.564335**	**0.870907**
Total features	**0.794589**	**0.800158**	**0.789021**	**0.791433**	**0.795736**	**0.589268**	**0.882882**

Table 13. Sub-predictor statistics for S-P class. Values in red font indicate the largest statistic value in each feature group.

Class S-nP	Accuracy	Sensitivity	Specificity	Precision	F1	MCC	AUROC
ZIMJ680104	0.697775	0.67005	0.7255	0.709437	0.689139	0.396203	0.762915
FASG760101	0.6173	0.63695	0.59765	0.612898	0.62467	0.234798	0.669534
GRAR740103	0.678525	0.6728	0.68425	0.68063	0.676664	0.357103	0.744073
RADA880106	0.6253	0.58055	0.67005	0.6376	0.607705	0.251627	0.672302
Vertical features	**0.791125**	**0.7627**	**0.81955**	**0.808718**	**0.785009**	**0.583229**	**0.873915**
GUYH850101	0.78475	0.8062	0.7633	0.773094	0.789258	0.570093	0.867598
MIYS990104	0.790275	0.8058	0.77475	0.781566	0.793458	0.580896	0.871936
PRAM900102	0.58555	0.46655	0.70455	0.612254	0.529459	0.176183	0.618341
PALJ810112	0.709825	0.72305	0.6966	0.704479	0.71362	0.419824	0.786046
ROBB760105	0.7535	0.75005	0.75695	0.755266	0.752613	0.507059	0.838519
PPIIPRO	0.597175	0.42545	0.7689	0.647984	0.513567	0.206941	0.632913
$\Delta G,N$	0.695025	0.7278	0.66225	0.683054	0.70469	0.390927	0.769178
$\Delta H_{ap},N$	0.546275	0.6325	0.46005	0.539487	0.582283	0.093961	0.57366
$\Delta H_{pol},N$	0.617125	0.69	0.54425	0.602246	0.643121	0.236801	0.661454
$T\Delta S_{conf},N$	0.652675	0.6847	0.62065	0.643589	0.663469	0.306011	0.708972
$\Delta\Delta G,N\text{-}D$	0.66655	0.67195	0.66115	0.664795	0.668329	0.333145	0.727326
$\Delta\Delta H_{ap},N\text{-}D$	0.69165	0.73315	0.65015	0.677023	0.703932	0.384673	0.764952
$\Delta\Delta H_{pol},N\text{-}D$	0.600525	0.6195	0.58155	0.596944	0.607967	0.201224	0.638616
$T\Delta\Delta S_{conf},N\text{-}D$	0.647875	0.6513	0.64445	0.646919	0.649078	0.295779	0.704685
Horizontal features	**0.818325**	**0.82745**	**0.8092**	**0.81277**	**0.819977**	**0.636863**	**0.898609**
Total features	**0.8376**	**0.8432**	**0.832**	**0.833922**	**0.838486**	**0.675324**	**0.918708**

Table 14. Sub-predictor statistics for S-nP class. Values in red font indicate the largest statistic value in each feature group.

Class T-P	Accuracy	Sensitivity	Specificity	Precision	F1	MCC	AUROC
ZIMJ680104	0.596562	0.551003	0.64212	0.60613	0.577107	0.193978	0.635436
FASG760101	0.608596	0.624642	0.59255	0.605623	0.614697	0.2175	0.660287
GRAR740103	0.634241	0.626074	0.642407	0.63686	0.63117	0.268733	0.682833
RADA880106	0.641404	0.59341	0.689398	0.656395	0.62315	0.28421	0.687897
Vertical features	**0.703295**	**0.709742**	**0.696848**	**0.700794**	**0.705197**	**0.406673**	**0.779973**
GUYH850101	0.691977	0.735817	0.648138	0.67625	0.704671	0.385647	0.766355
MIYS990104	0.704155	0.743266	0.665043	0.689275	0.715105	0.409807	0.775734
PRAM900102	0.600573	0.497708	0.703438	0.626777	0.554658	0.205645	0.631605
PALJ810112	0.632378	0.689685	0.575072	0.619099	0.652177	0.266828	0.684848
ROBB760105	0.659456	0.712034	0.606877	0.64435	0.676375	0.320855	0.711871
PPIIPRO	0.640258	0.513467	0.767049	0.687936	0.587811	0.290073	0.702787
ΔG,N	0.637536	0.667622	0.60745	0.629882	0.647905	0.275879	0.695882
ΔH_{ap},N	0.582665	0.470774	0.694556	0.606579	0.529857	0.169718	0.613442
ΔH_{pol},N	0.603295	0.658453	0.548138	0.593015	0.623854	0.208042	0.643289
$T\Delta S_{conf}$,N	0.574355	0.659026	0.489685	0.563547	0.607445	0.151014	0.60039
ΔΔG,N-D	0.64298	0.607163	0.678797	0.653712	0.629321	0.286842	0.701574
$\Delta\Delta H_{ap}$,N-D	0.620917	0.661032	0.580802	0.612181	0.635283	0.242993	0.67731
$\Delta\Delta H_{pol}$,N-D	0.602579	0.537249	0.667908	0.61822	0.574768	0.207032	0.649613
$T\Delta\Delta S_{conf}$,N-D	0.633954	0.575358	0.69255	0.651677	0.610863	0.269926	0.684493
Horizontal features	**0.723782**	**0.760172**	**0.687393**	**0.708504**	**0.733173**	**0.449166**	**0.80064**
Total features	**0.741404**	**0.767908**	**0.7149**	**0.729282**	**0.747969**	**0.483678**	**0.8199**

Table 15. Sub-predictor statistics for T-P class. Values in red font indicate the largest statistic value in each feature group.

Class T-nP	Accuracy	Sensitivity	Specificity	Precision	F1	MCC	AUROC
ZIMJ680104	0.617098	0.583364	0.650832	0.625888	0.60373	0.234852	0.660006
FASG760101	0.57403	0.609057	0.539002	0.569071	0.588254	0.148527	0.602642
GRAR740103	0.619131	0.607763	0.630499	0.621895	0.61446	0.23854	0.663905
RADA880106	0.586784	0.521442	0.652126	0.600085	0.557798	0.175196	0.618641
Vertical features	**0.697782**	**0.686322**	**0.709242**	**0.702684**	**0.694134**	**0.39596**	**0.767477**
GUYH850101	0.699723	0.729575	0.669871	0.688637	0.708367	0.400345	0.770304
MIYS990104	0.703974	0.721996	0.685952	0.69702	0.709065	0.408498	0.777033
PRAM900102	0.563863	0.431608	0.696118	0.586981	0.497185	0.132533	0.58646
PALJ810112	0.646026	0.655638	0.636414	0.643223	0.649048	0.292422	0.70165
ROBB760105	0.67366	0.674492	0.672828	0.673373	0.673683	0.347582	0.737826
PPIIPRO	0.584196	0.402403	0.765989	0.632722	0.491213	0.181002	0.622557
$\Delta G,N$	0.62597	0.653974	0.597967	0.619301	0.635956	0.252567	0.673932
$\Delta H_{ap},N$	0.522089	0.479667	0.56451	0.524836	0.500084	0.044667	0.536126
$\Delta H_{pol},N$	0.576617	0.61756	0.535675	0.570771	0.593159	0.15382	0.60663
$T\Delta S_{conf},N$	0.597135	0.642514	0.551756	0.589062	0.614575	0.195121	0.642398
$\Delta\Delta G,N\text{-}D$	0.6122	0.573937	0.650462	0.621622	0.596631	0.225185	0.651545
$\Delta\Delta H_{ap},N\text{-}D$	0.619501	0.657671	0.581331	0.610992	0.633263	0.239947	0.675139
$\Delta\Delta H_{pol},N\text{-}D$	0.574861	0.553604	0.596118	0.578273	0.565429	0.149968	0.600651
$T\Delta\Delta S_{conf},N\text{-}D$	0.599168	0.575231	0.623105	0.604164	0.589203	0.19864	0.632708
Horizontal features	**0.716636**	**0.715896**	**0.717375**	**0.717161**	**0.716156**	**0.433723**	**0.792609**
Total features	**0.729945**	**0.735305**	**0.724584**	**0.727746**	**0.731189**	**0.460337**	**0.81032**

Table 16. Sub-predictor statistics for T-nP class. Values in red font indicate the largest statistic value in each feature group.

Class Tyr	Accuracy	Sensitivity	Specificity	Precision	F1	MCC	AUROC
ZIMJ680104	0.619481	0.602857	0.636104	0.623745	0.612989	0.239191	0.665827
FASG760101	0.577662	0.616104	0.539221	0.572463	0.593164	0.155954	0.609484
GRAR740103	0.615325	0.604675	0.625974	0.618149	0.611077	0.230897	0.67007
RADA880106	0.593636	0.549091	0.638182	0.60318	0.574597	0.188191	0.637699
Vertical features	**0.678052**	**0.643377**	**0.712727**	**0.691855**	**0.666293**	0.357387	**0.741669**
GUYH850101	0.681429	0.746753	0.616104	0.660855	0.701002	0.366175	0.74083
MIYS990104	0.692078	0.741299	0.642857	0.675707	0.706663	0.386327	0.756548
PRAM900102	0.554545	0.439221	0.66987	0.571021	0.496193	0.1122	0.577257
PALJ810112	0.628701	0.658961	0.598442	0.621619	0.639436	0.258134	0.678237
ROBB760105	0.648442	0.681039	0.615844	0.640043	0.659655	0.297698	0.705415
PPIIPRO	0.585065	0.414545	0.755584	0.629146	0.49953	0.181036	0.619905
ΔG,N	0.607143	0.661039	0.553247	0.596803	0.627118	0.21569	0.659779
ΔH_{ap},N	0.528701	0.406494	0.650909	0.538099	0.462391	0.059241	0.529689
ΔH_{pol},N	0.574026	0.605195	0.542857	0.569747	0.586784	0.148431	0.600974
$T\Delta S_{conf}$,N	0.57	0.603377	0.536623	0.565604	0.583715	0.14041	0.595487
ΔΔG,N-D	0.581688	0.592208	0.571169	0.580186	0.585936	0.163518	0.622842
$\Delta\Delta H_{ap}$,N-D	0.613247	0.672208	0.554286	0.601535	0.634806	0.228138	0.657273
$\Delta\Delta H_{pol}$,N-D	0.550649	0.560519	0.540779	0.549677	0.554867	0.101372	0.581004
$T\Delta\Delta S_{conf}$,N-D	0.591169	0.605714	0.576623	0.58918	0.597043	0.182583	0.630948
Horizontal features	**0.694675**	**0.713247**	**0.676104**	**0.688193**	**0.700174**	**0.390005**	**0.76217**
Total features	**0.718442**	**0.717143**	**0.71974**	**0.719261**	**0.717993**	**0.43715**	**0.791034**

Table 17. Sub-predictor statistics for Y class. Values in red font indicate the largest statistic value in each feature group.

Class S-P	Accuracy	Sensitivity	Specificity	Precision	F1	MCC	AUROC
Vertical features	0.752725	0.738784	0.766667	0.760029	0.749216	0.505698	0.835519
Horizontal features	0.78207	0.792733	0.771406	0.77631	0.784389	0.564335	0.870907
Total features	0.794589	0.800158	0.789021	0.791433	0.795736	0.589268	0.882882
Class S-nP	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1**	**MCC**	**AUROC**
Vertical features	0.791125	0.7627	0.81955	0.808718	0.785009	0.583229	0.873915
Horizontal features	0.818325	0.82745	0.8092	0.81277	0.819977	0.636863	0.898609
Total features	0.8376	0.8432	0.832	0.833922	0.838486	0.675324	0.918708
Class T-P	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1**	**MCC**	**AUROC**
Vertical features	0.703295	0.709742	0.696848	0.700794	0.705197	0.406673	0.779973
Horizontal features	0.723782	0.760172	0.687393	0.708504	0.733173	0.449166	0.80064
Total features	0.741404	0.767908	0.7149	0.729282	0.747969	0.483678	0.8199
Class T-nP	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1**	**MCC**	**AUROC**
Vertical features	0.697782	0.686322	0.709242	0.702684	0.694134	0.39596	0.767477
Horizontal features	0.716636	0.715896	0.717375	0.717161	0.716156	0.433723	0.792609
Total features	0.729945	0.735305	0.724584	0.727746	0.731189	0.460337	0.81032
Class Y	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1**	**MCC**	**AUROC**
Vertical features	0.678052	0.643377	0.712727	0.691855	0.666293	0.357387	0.741669
Horizontal features	0.694675	0.713247	0.676104	0.688193	0.700174	0.390005	0.76217
Total features	0.718442	0.717143	0.71974	0.719261	0.717993	0.43715	0.791034

Table 18. Full PHOSforUS predictor performances calculated from X10 cross-validation.

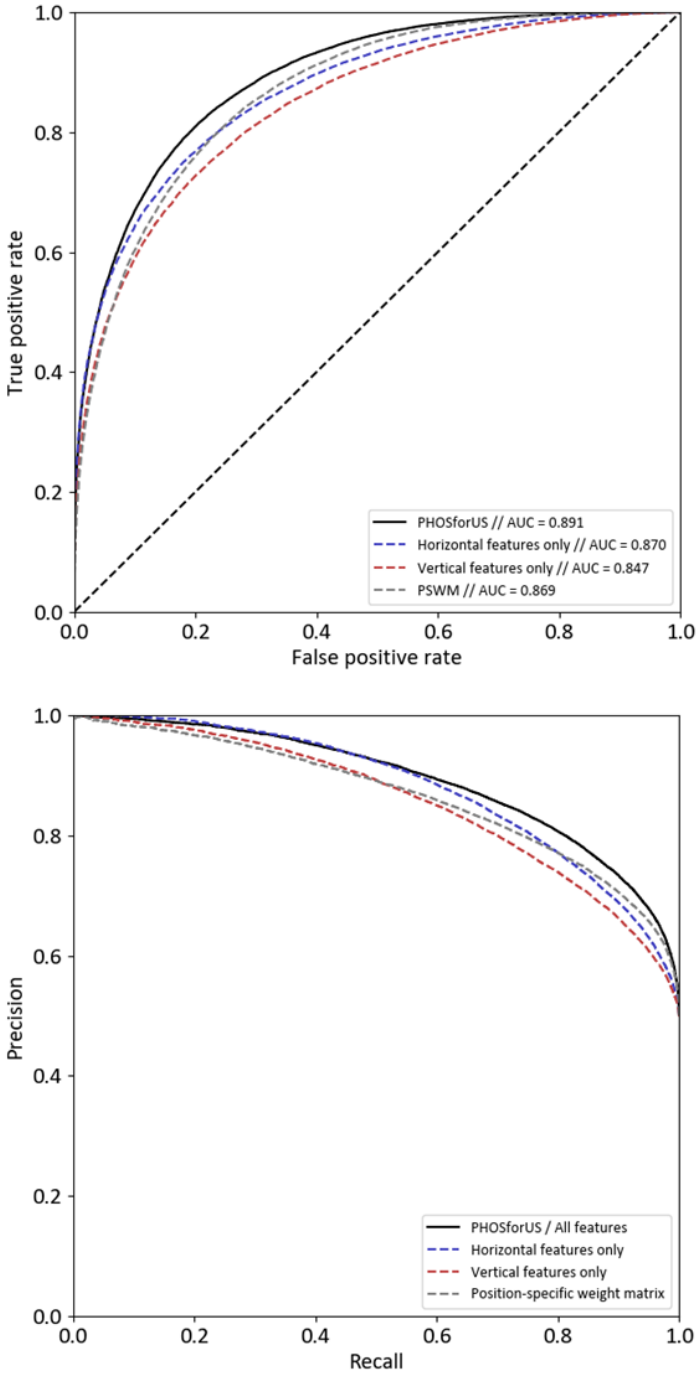

Figure 55. Receiver-operating characteristics (ROC) curve (upper panel) and precision-recall curve (lower panel) of PHOSforUS predictor / sub-predictors along with PSWM-based prediction results

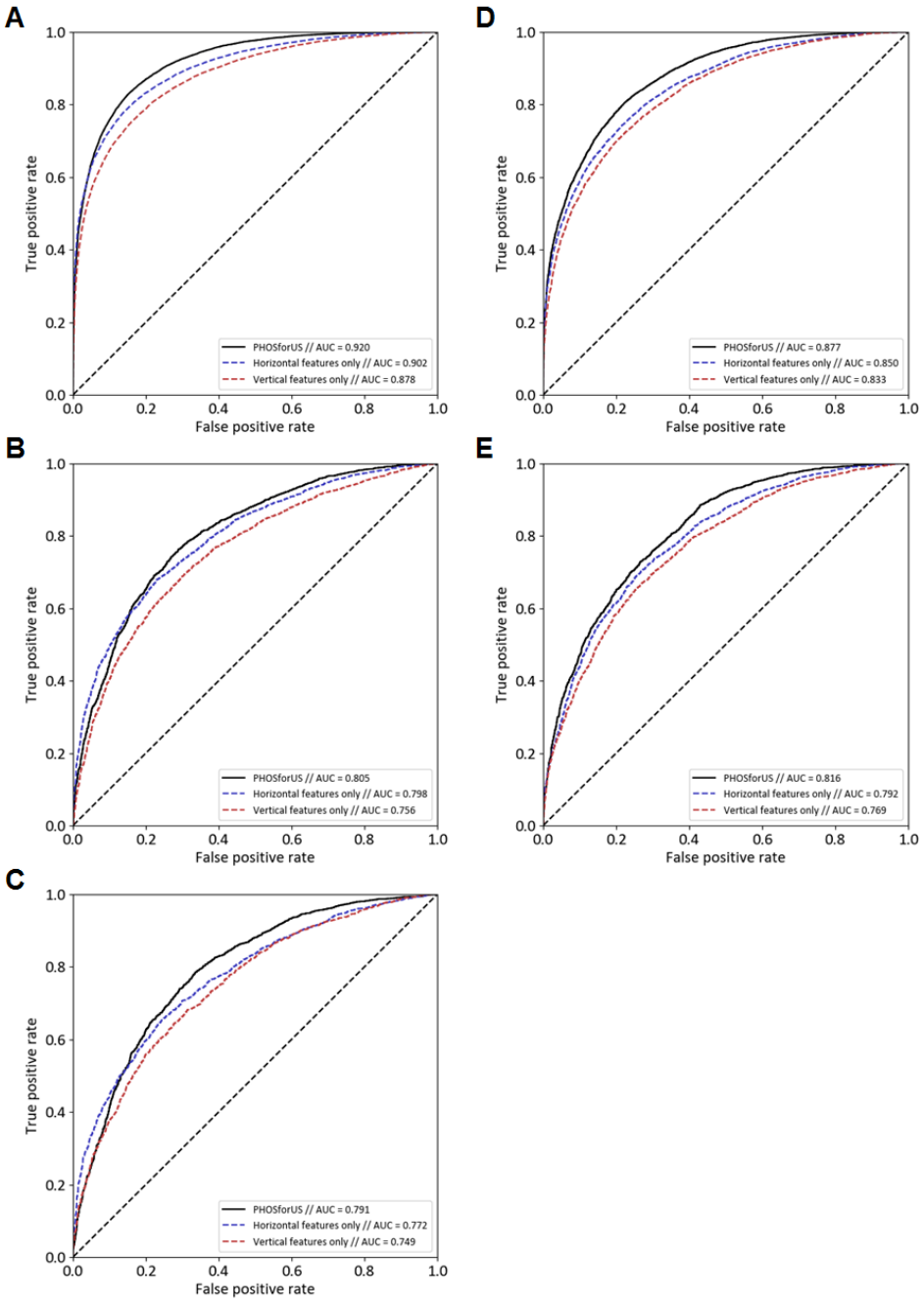

Figure 56. Subclass-specific ROC curves of PHOSforUS constituent predictors. A. S-nP sites. B. T-nP sites. C. Tyrosine sites. D. S-P sites. E. T-P sites.

more informative in identifying S-P and T-P sites. This result is largely consistent with previously identified class-specific association of biophysical properties with phosphorylation sites (section 3.3.1).

Evaluation of the downstream gradient boosting classifiers which take pre-processed values from aforementioned sub-predictors with 10x cross-validation demonstrated that all five phosphorylation site classes could be reasonably predicted from biophysical property values (Table 18). The partial predictors trained with either vertical or horizontal information showed comparable predictive power (Tables 13-17), with the horizontal combination, which includes thermodynamic information calculated with COREX/eSCAPE, showing the best performance.

While utilizing both vertical and horizontal information results in improved predictive performance (Figure 55, black curves), horizontal information was generally more effective (as measured by AUROC) across phosphorylation subclasses than vertical information (Figures 55, 56). These results suggest that conformational equilibrium would be a predominant factor which affects protein phosphorylation.

-4.3.4. Comparative analysis

Among several dozen phosphorylation site predictors currently available, six tools – Disphos (102), Musite (170), Netphos3.1 (38), PhosphoSVM (171), PhosPred-RF (172) and RF-phos (173) - were selected and compared with PHOSforUS for the objective assessment of predictive performances. These tools were chosen for high accessibility and ability to handle large datasets utilized for comparison.

The ROC curves indicated the methods could be divided into two groups, with more effective group included tools which incorporated predicted IDR information (Table 19). For all phosphorylation site classes, PHOSforUS showed the highest AUROC and MCC values (Figure 57) (Table 19). Due to the possibility that phosphorylation sites in the testing set were not already contained in the training sets used for other predictors, it is probable that the improvement of predictive performance shown here is a conservative estimate.

Class S-nP	Accuracy	Sensitivity	Specificity	Precision	F1 score	MCC	AUROC
PHOSforUS	0.795	0.74	0.85	0.834396	0.782946	0.595424	0.8707
Disphos	0.717	0.544	0.89	0.833368	0.657429	0.463373	0.82301
Musite	0.669	0.448	0.89	0.804146	0.574517	0.377379	0.78346
Netphos3.1	0.616	0.862	0.37	0.57799	0.691847	0.266571	0.71711
Rfphos	0.637	0.372	0.902	0.791095	0.504715	0.322944	0.74387
PhosPred-RF	0.772	0.654	0.89	0.857754	0.740594	0.561035	0.81251
PhosphoSVM	0.656	0.366	0.946	0.873751	0.515124	0.383791	0.81356
Class T-nP	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1 score**	**MCC**	**AUROC**
PHOSforUS	0.687	0.64	0.734	0.706602	0.671354	0.375909	0.74322
Disphos	0.578	0.37	0.786	0.632685	0.466229	0.171279	0.62811
Musite	0.599	0.366	0.832	0.684903	0.475075	0.223637	0.67413
Netphos3.1	0.531	0.598	0.464	0.528168	0.560629	0.062365	0.53124
Rfphos	0.61	0.372	0.848	0.711223	0.486631	0.25072	0.67351
PhosPred-RF	0.666	0.578	0.754	0.701385	0.633465	0.337389	0.71469
PhosphoSVM	0.603	0.288	0.918	0.779423	0.419856	0.265534	0.72002
Class Y	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1 score**	**MCC**	**AUROC**
PHOSforUS	0.663	0.588	0.738	0.693753	0.634787	0.331058	0.72352
Disphos	0.595	0.412	0.778	0.653431	0.504334	0.2056	0.65703
Musite	0.6	0.578	0.622	0.606359	0.590998	0.200748	0.65107
Netphos3.1	0.603	0.55	0.656	0.616933	0.580175	0.208103	0.62247
Rfphos	0.594	0.476	0.712	0.62333	0.539352	0.193675	0.63655
PhosPred-RF	0.62	0.744	0.496	0.596424	0.662059	0.247556	0.68372
PhosphoSVM	0.619	0.686	0.552	0.604983	0.642854	0.240269	0.67874
Class S-P	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1 score**	**MCC**	**AUROC**
PHOSforUS	0.763	0.72	0.806	0.787563	0.752005	0.528191	0.84546
Disphos	0.69	0.692	0.688	0.691998	0.691489	0.380551	0.75849
Musite	0.631	0.868	0.394	0.591041	0.702042	0.297662	0.71465
Netphos3.1	0.532	0.972	0.092	0.517175	0.67509	0.130089	0.63346
Rfphos	0.608	0.836	0.38	0.57441	0.680826	0.242584	0.67445
PhosPred-RF	0.66	0.95	0.37	0.602202	0.73677	0.39284	0.73841
PhosphoSVM	0.553	0.98	0.126	0.528897	0.686878	0.201449	0.70333
Class T-P	**Accuracy**	**Sensitivity**	**Specificity**	**Precision**	**F1 score**	**MCC**	**AUROC**
PHOSforUS	0.69	0.666	0.714	0.700031	0.682284	0.380759	0.76799
Disphos	0.592	0.762	0.422	0.568356	0.650632	0.197605	0.66303
Musite	0.597	0.838	0.356	0.565398	0.675205	0.221692	0.6354
Netphos3.1	0.52	0.95	0.09	0.510748	0.664287	0.079128	0.59619
Rfphos	0.583	0.87	0.296	0.552847	0.675764	0.204821	0.6442
PhosPred-RF	0.592	0.956	0.228	0.553552	0.700941	0.269366	0.68277
PhosphoSVM	0.59	0.956	0.224	0.551987	0.699722	0.26752	0.69151

Table 19. Full comparative analysis data of PHOSforUS with current phosphorylation site predictors

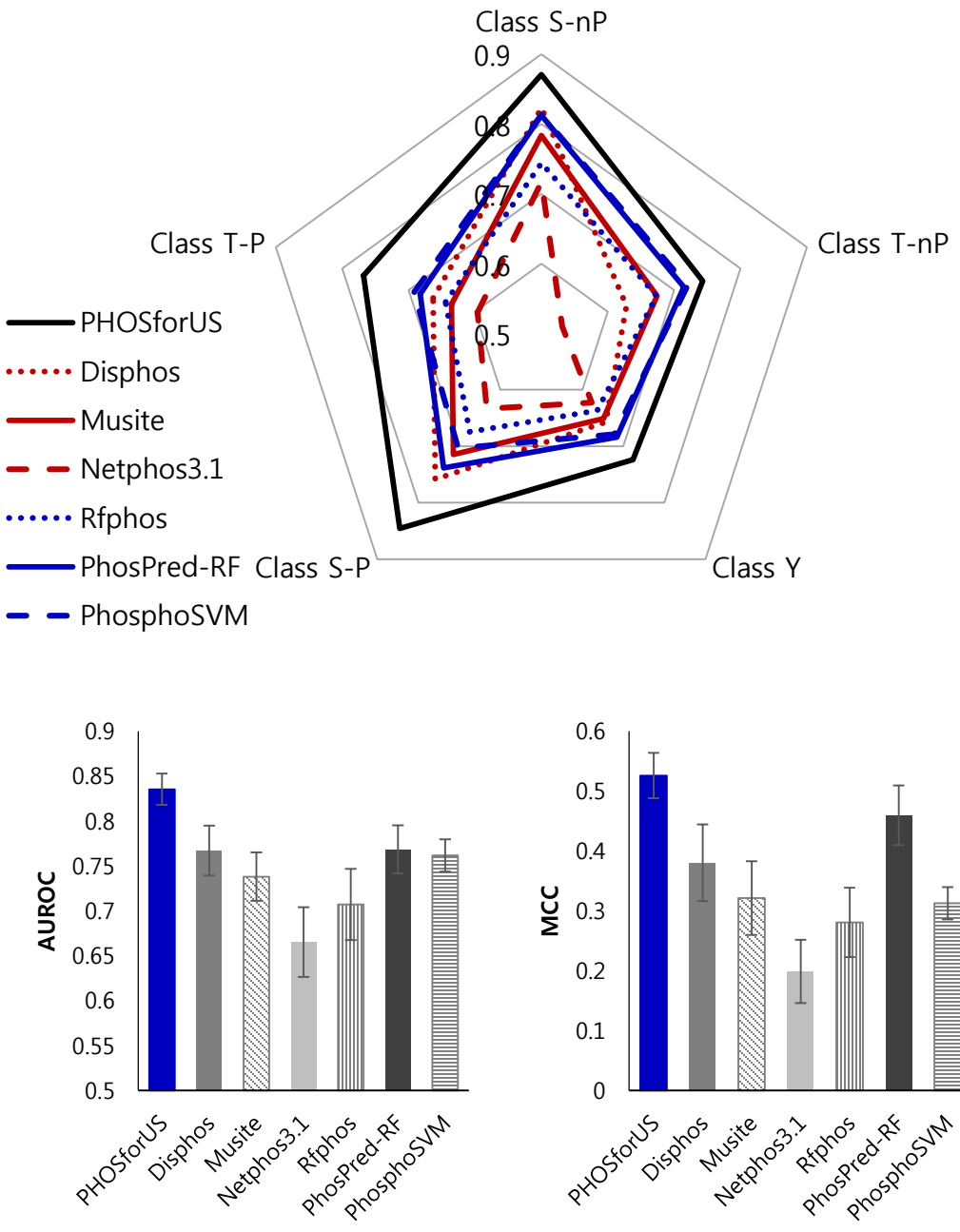

Figure 57. Comparative effectiveness of protein phosphorylation site prediction by PHOSforUS Upper panel: Comparative effectiveness of protein phosphorylation site prediction by PHOSforUS. For five classes of phosphorylation site, PHOSforUS AUROC values meet or exceed those obtained on the identical data with six existing prediction tools. Lower panels: Weighted averages of AUROC (lower left) and MCC (lower right) again suggest PHOSforUS has a superior predictive performance overall.

[4-4] Discussion

All these results suggest that the direct consideration of horizontal information identified to be associated with protein phosphorylation and separation of serine and threonine phosphorylation sites based on +1 residue identity could be utilized to design a powerful phosphorylation site predictor. Even without direct consideration of sequence information and similarity annotation, PHOSforUS resulted in better predictive performances than other existing prediction tools.

We devised a schematic which represents phosphorylation reaction as two distinct processes, which are associated with different biophysical properties. The 'vertical' information is more reflective of classical, structure-based perspective of proteins, where the conserved site-specific sequence elements provide a framework needed for specific binding. In effect, the magnitude of conservation could serve as a proxy for the energy of the interaction, which is consistent with known similarity in theoretical and experimental energy changes analyzed within folded proteins (174).

However, the unique feature of the approach here is the explicit consideration of sequence-averaged / horizontal information which encodes conformational thermodynamics embedded in the sequence. It is important that both information types could be encoded by protein sequence and conserved in the multiple sequence alignment of substrate (Figure 4-8), with a pivotal difference being that the horizontal information tend to be more dispersed – makes it harder to find via traditional alignment tools (175). This could be an implication of different strategy of evolution which allows rapid testing of functional amino acid mutations within conserved IDRs. Support for the relevance of horizontal information comes from the direct comparison of predictor statistics, such as accuracy, AUROC and MCC, which reveals that horizontal properties outperform vertical properties in every phosphorylation subclass (148) (Figure 56, Tables 13-17).

Another key feature of PHOSforUS which improved its predictive performances was the explicit consideration of the presence of +1 proline. As it is elaborated in previous chapters, +1 proline is an indicator of distinct biophysical nature, which justifies dividing serine/threonine phosphorylation sites into two groups respectively and analyzing each group separately. Albeit reduced number of training samples could have increased the chance of overfitting, utilizing Naïve Bayes classifier and nested algorithm architecture allowed

to exploit class-specific information while avoiding predictors being overfitted. In addition, considering the biophysical properties associated with S/T-P sites - PII propensity, C'-terminal alpha helix propensity and thermodynamic descriptors for instance - are mostly horizontal in nature, our approach would worked better than other predictors which does not take account of this type of information.

Again, phosphorylation site prediction is one of possible way of utilizing this thermodynamic framework. Coupled with different types of information, the underlying concept of PHOSforUS could be implemented to predict different biochemical reaction involving IDRs, such as other PTMs or protein-protein interactions. This could be an alternative mode of assessing characteristics of unknown protein in high-throughput manner, which would be complementary to other predictors based on sequence information and other types of annotations.

[5] +1 prolines of S/T-P phosphorylation sites are evolutionarily conserved

[5-1] Introduction

In this chapter, I'd like to address is the evolutionary history of S/T-P sites and suggest possible explanation why S/T-P phosphorylation sites become abundant in human phosphoproteome.

I already demonstrated that S/T-P classes have been enriched along with accumulation of IDRs in eukaryotes, and it seems to be associated with average number of disordered region per each protein species. Interestingly, phosphorylation sites and IDRs have the opposite influences on sequence conservation: while there is a doubt on the significance of conservation (176), as one of the functional sites in proteins, phosphorylation sites are known to be evolutionarily conserved across the ortholog proteins (177). On the other hand, IDRs are known for higher tolerance to mutations, including single-site mutations, small indels and frameshifts, allowing ortholog sequences to be highly divergent (178). This brings a question how the phosphorylation classes (S/T-P, other S/T & tyrosine) conserved differently across the orthologs.

In addition, if the S/T-P site enrichment is this high - more than 30% - I suspected that the corresponding 'traces' should be found in evolutionary record, such as ortholog sequences of human phosphoprotein. There are couple of possible cases for these traces: first, nascent (with younger origin) phosphorylation sites might have higher frequency of being S/T with P sites. Also, +1 proline might be more conserved than other amino acids at +1 site relative to phosphorylation site. Another possibility would be the indels, occurring in much higher frequency in IDRs, produce higher amount of SP/TP dipeptides, which potentially be recognized as a valid substrate of kinase and phosphorylated.

So I used ancestral sequence reconstruction based on ortholog sequences of known phosphoproteins from human and mouse to recreate evolutionary records of individual phosphorylation sites. This revealed different classes of phosphorylation sites emerge in different rate: S/T-P sites are likely to have more recent origin than other S/T phosphorylation sites and tyrosine phosphorylation sites. Also, amino acid substitution rates found in +1 sites were significantly different from those found on other sites nearby phosphorylation sites or non-

phosphorylated sequences, which support gradual accumulation of +1 proline nearby S/T phosphorylation sites.

Moreover, there were two unexpected discoveries. First, +1 proline residues are more likely to predate actually phosphorylated S/T residues in the evolutionary records. Second, substitution rates between phosphorylated serine and threonine residues were remarkably different for S/T-nP sites than those found in non-phosphorylated sites or S/T-P sites. Coupled with the biophysical characteristics of S/T-P sites identified in previous chapter, it provided me an interesting hypotheses about fundamental properties of S/T-P phosphorylation.

[5-2] Approaches

-5.2.1. Data sources

I used the same human phosphorylation site datasets used for chapter 2 and 3. Mouse phosphorylation site datasets were similarly generated with SWISS-PROT annotations (25) and low-throughput (LTP) subset of PhosphoSitePlus (35).

Ortholog annotations were collected from OMA (150) and EggNOG (179) databases independently. Reference proteome sequences and site-specific functional annotation were retrieved from Uniprot (25). I limited the range of species to mammalian level not only to keep the quality of multiple sequence alignment but also to ensure that there is no substantial difference between kinase sets. List of species included is shown in table 21. Phylogenetic relationship of these species is visualized as a tree in Figure 58. Total of 6.904 ortholog groups which includes one of human phosphoprotein were reconstructed from member labels from ortholog databases and protein sequences from reference proteomes. Resulting statistics of ortholog groups are shown in Table 20.

Class	Phos-site. Local align.	Average number of orthologs	Associated phosphoproteins	Nonphos. Local align.
S-P	8736	31.33	3553	24742
S-nP	18774	31.41	5437	388214
T-P	2308	31.62	1419	16724
T-nP	2662	30.86	1746	245983
Y	1751	31.21	937	123657
Total / Mean	34231	31.35074	6904	799320

Table 20. Statistics of phosphoprotein ortholog dataset

Abbr.	Scientific name	Common name	Abbr.	Scientific name	Common name
HUMAN	Homo sapiens	Human	**DIPOR**	Dipodomys ordii	Ord's kangaroo rat
PANPA	Pan paniscus	Bonobo	**FUKDA**	Fukomys damarensis	Damaraland mole-rat
PANTR	Pan troglodytes	Chimpanzee	**HETGA**	Heterocephalus glaber	Naked mole-rat
GORGO	Gorilla gorilla	Western gorilla	**CAVPO**	Cavia porcellus	Guinea pig
PONAB	Pongo abelii	Sumatran orangutan	**ICTTR**	Ictidomys tridecemlineatus	Thirteen-lined ground squirrel
NOMLE	Nomascus leucogenys	Northern white-cheeked gibbon	**RABIT**	Oryctolagus cuniculus	European rabbit
MACFA	Macaca fascicularis	Crab-eating macaque	**URSAM**	Ursus americanus	American black bear
MACMU	Macaca mulatta	Rhesus macaque	**URSMA**	Ursus maritimus	Polar bear
MACNE	Macaca nemestrina	Southern pig-tailed macaque	**AILME**	Ailuropoda melanoleuca	Giant panda
CERAT	Cercocebus atys	Sooty mangabey	**MUSPF**	Mustela putorius furo	Ferret
MANLE	Mandrillus leucophaeus	Drill	**CANLF**	Canis lupus familiaris	Dog
PAPAN	Papio anubis	Olive baboon	**VULVU**	Vulpes vulpes	Red fox
CHLSB	Chlorocebus sabaeus	Green monkey	**FELCA**	Felis catus	Cat
RHIBE	Rhinopithecus bieti	Black snub-nosed monkey	**BOVIN**	Bos taurus	Cattle
RHIRO	Rhinopithecus roxellana	Golden snub-nosed monkey	**SHEEP**	Ovis aries	Sheep
COLAP	Colobus angolensis palliatus	Peter's angola colobus	**TURTR**	Tursiops truncatus	Common bottlenose dolphin
CALJA	Callithrix jacchus	Common marmoset	**PIGXX**	Sus scrofa	Pig
AOTNA	Aotus nancymaae	Nancy ma's night monkey	**HORSE**	Equus caballus	Horse
SAIBB	Saimiri boliviensis boliviensis	Bolivian squirrel monkey	**MYOLU**	Myotis lucifugus	Little brown bat
TARSY	Carlito syrichta	Philippine tarsier	**ERIEU**	Erinaceus europaeus	European hedgehog
OTOGA	Otolemur garnettii	Northern greater galago	**LOXAF**	Loxodonta africana	African bush elephant
PROCO	Propithecus coquereli	Coquerel's sifaka	**MONDO**	Monodelphis domestica	Gray short-tailed opossum
MOUSE	Mus musculus	House mouse	**SARHA**	Sarcophilus harrisii	Tasmanian devil
RATNO	Rattus norvegicus	Brown rat	**ORNAN**	Ornithorhynchus anatinus	Platypus
CRIGR	Cricetulus griseus	Chinese hamster			

Table 21. List of species included in ortholog dataset

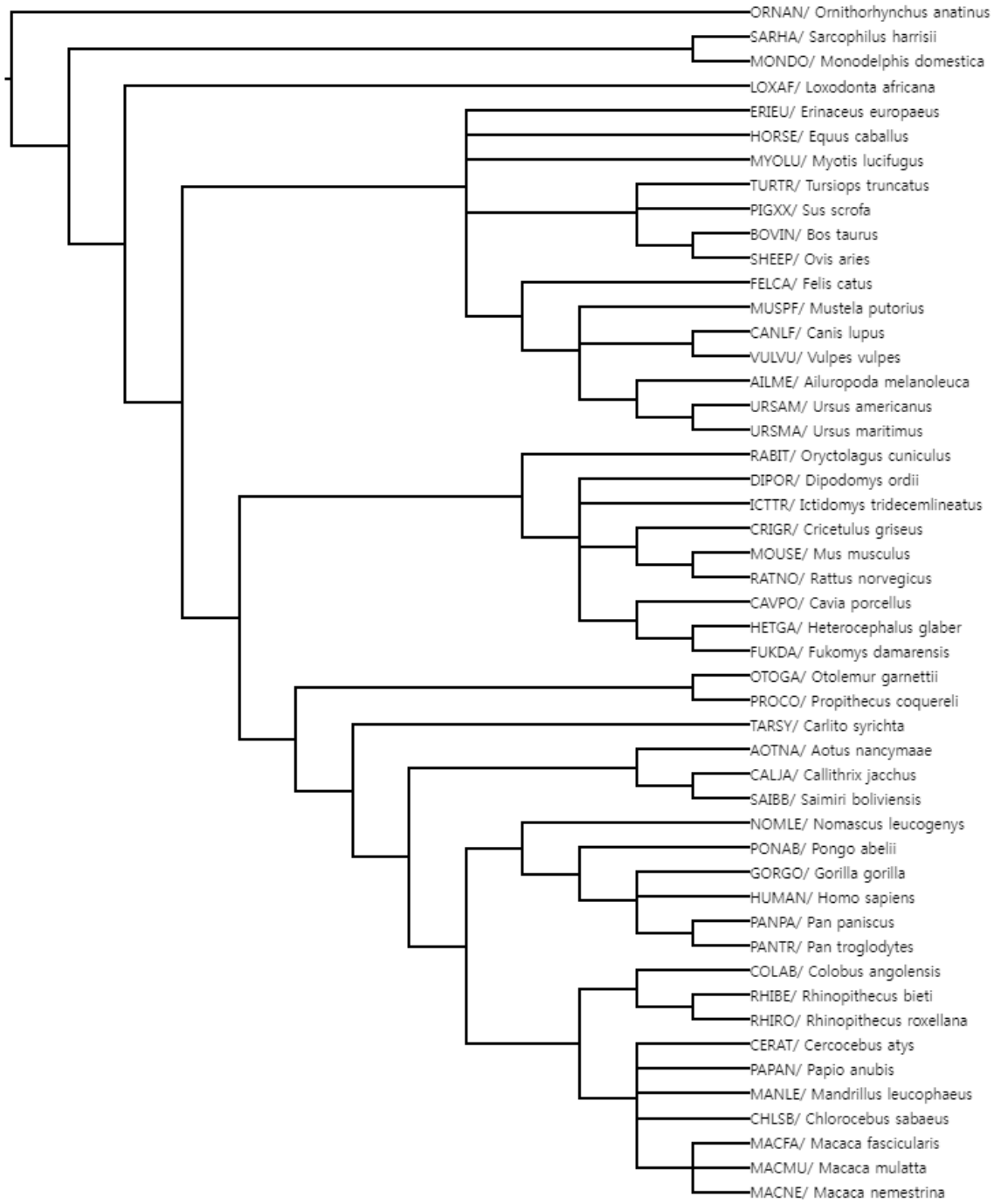

Figure 58. Phylogenetic relationship of analyzed species

-5.2.2. Multiple sequence alignment

Multiple sequence alignment is done with ClustalW alignment tool (180). From the resulting alignments, I searched for the alignment slice (example showed in figure 59) with S / T / Y residues (phosphorylation sites included) in either human or mouse sequence. Local alignment for +-5 region around the identified slices were retrieved. Site-specific conservation rates were calculated for each amino acids accordingly. Total numbers of alignments with / without phosphorylation sites at the center were 34,231 and 799,320 respectively (Table 20).

-5.2.3. Ancestral sequence reconstruction

Hierarchical ancestral sequence reconstruction based on phylogenetic information is done with maximum likelihood-based reconstruction algorithm (181). For each site, the conditional probability of having an ancestral state A given observed states B = {B_1, B_2, …, B_n} could be written as a sum of conditional probabilities of having ancestral codons A = {A_1, A_2, …, A_m} given observed states..

$$\mathrm{P}(\mathrm{A}|B_1, B_2, \cdots, B_n) = \sum_{j=1}^{m} \mathrm{P}\left(A_j \middle| B_1, B_2, \cdots, B_n\right) \quad \ldots (33)$$

Applying Bayes' theorem with the assumption of independence between observed states transforms the expression as:

$$\mathrm{P}(\mathrm{A}|B_1, B_2, \cdots, B_n) \propto \sum_{j=1}^{m} \mathrm{P}(A_j) \prod_{i=1}^{n} P(B_i|A_j) \quad \ldots (34)$$

Again, probability of having a particular observed state B_i is equal to the sum of probabilities of having corresponding codons B_i = {B_{i1}, B_{i2}, …, B_{io}}. This allows to calculate conditional probability of an ancestral state from codon usage and codon substitution frequency:

$$\mathrm{P}(\mathrm{A}|B_1, B_2, \cdots, B_n) \propto \sum_{j=1}^{m} \mathrm{P}(A_j) \prod_{i=1}^{n} \sum_{h=1}^{o} P(B_{ih}|A_j) \quad \ldots (35)$$

Here, $P(A_j)$ is a codon usage frequency of A_j and $P(B_{ih}|A_j)$ is a frequency of substitution from A_j to B_{ih}, which were retrieved from previous studies (182, 183). Based on these parameters, reconstruction algorithm selects

the ancestral state which maximizes the conditional probability given observed states.

$$\hat{y} = \underset{K\in\{A,C,\ldots,\mathrm{Y},-\}}{\operatorname{argmax}} \sum_{j=1}^{m} \mathrm{P}(A_{Kj}) \prod_{i=1}^{n} \sum_{h=1}^{o} P(B_{ih}|A_{Kj}) \quad \ldots (36)$$

Ancestral sequence is reconstructed for every nodes in the phylogenetic tree (figure x), in hierarchical manner until the tree root is reached. The algorithm assumes Markov property - only the information from directly connected nodes is utilized as an input for each reconstruction step.

-5.2.4. Rate of evolution

Rate of evolution for site m was calculated from the observed mutation rates for node N as follows (184):

$$r_{ev} = \frac{f_{mut,m}(A_N, RS_N)}{t_{div}(N)} = \frac{\sum_{i=1}^{h} P(s_{i,m} != RS_{N,m})}{t_{div}(N) \cdot h} \quad \ldots (37)$$

Here, $A_N = \{s_1, s_2, \ldots, s_h\}$ is a sequence alignment containing all sequences associated with node N, RS_N is a reconstructed ancestral sequence using alignment, and t_{div} is a divergence time of node N. Using this formula, it was able to calculate rates of evolution for the same site but with reference nodes with different divergence time, which in turn allowed statistical analysis of differences between different phosphorylation site classes and between phosphorylation site and non-phosphorylated sequences.

[5-3] Results

-5.3.1. Individual S/T-P phosphorylation sites are more likely to have younger origin

I identified phosphorylation sites which could not be phosphorylated in reconstructed ancestral sequences, which I classified as nascent phosphorylation sites. The times of origin of each nascent phosphorylation sites were estimated from the known divergence dates of mammalian clades (185, 186). 12,657 out of 34,231 phosphorylation sites (37.0%) were labeled as nascent phosphorylation sites.

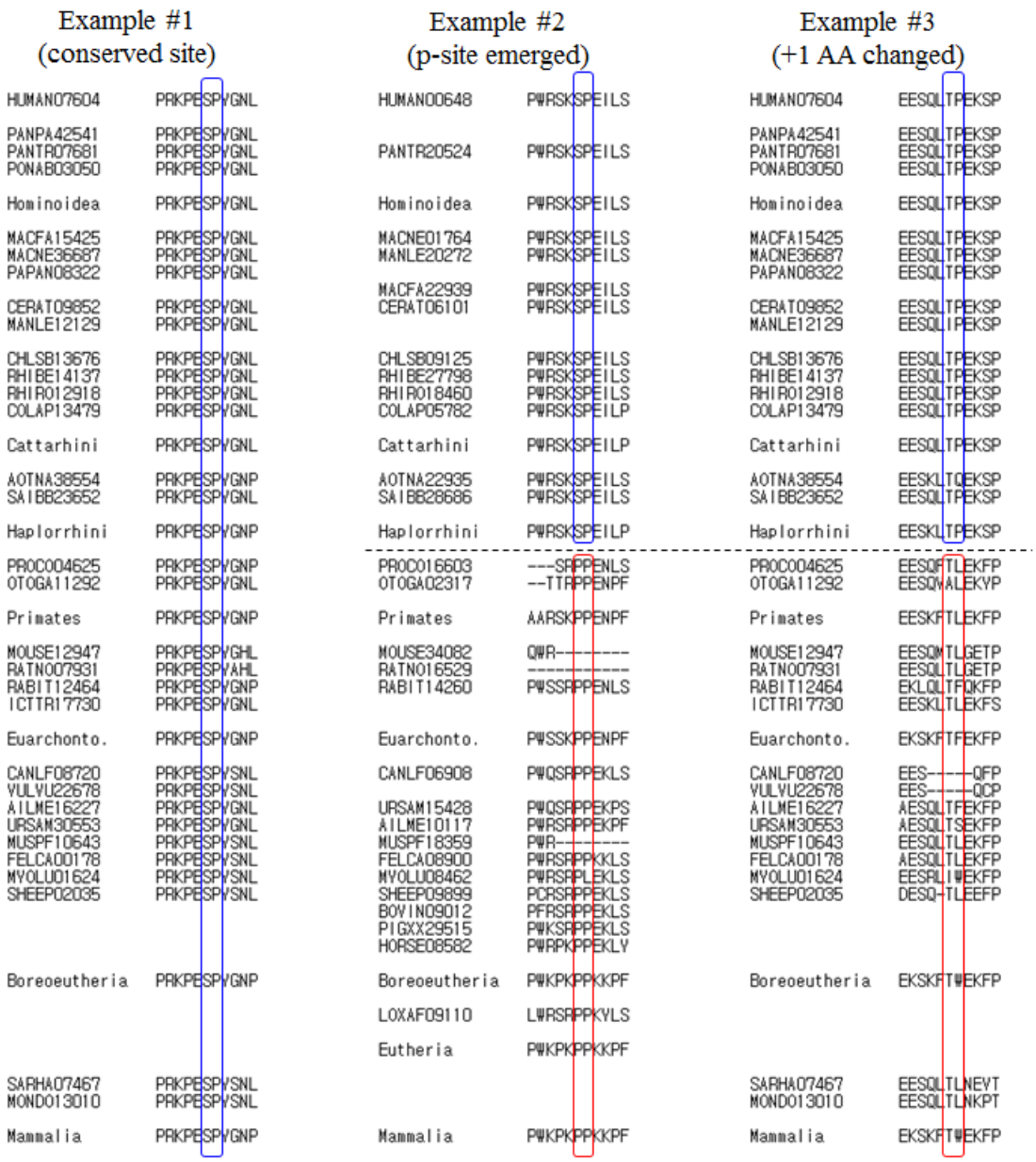

Figure 59. Example ortholog local alignments

I'd like to show three examples of local alignments around phosphorylation sites, which show different evolutionary patterns across orthologs (Figure 59). First example shows a phosphorylation site which is conserved in all mammalian species, indicating this site emerged before monotremes and therians diverged (>210~165 Mya). Second example shows a phosphorylation sites which is found in all primate species but not in other mammalian sequences, implying the time of origin of this phosphorylation site is after primates emerge from the common ancestor of euarchontoglires (75 ~ 85 Mya) but before haplorrhine and strepsirrhine (~65 Mya) are separated. The third example shows the case which the phosphorylated S/T residue is conserved among the orthologs but +1 proline is only found in haplorrhines. Considering the specificity of kinases to +1 residues (Section 2.3.2), changing proline on +1 site to other amino acid or vice versa would be likely to change the status as a phosphorylation site.

S/T-P sites were significantly more likely to be found among nascent phosphorylation sites (Figure 60). 38.4% of phosphorylation sites which emerged in placentalia clade (<~100 Mya) were S/T-P sites. Further classification of nascent phosphorylation sites with time of origin reveals more interesting pattern (Figure 61): the frequency of S/T-P sites was the highest (44.6%) for the 'youngest' group, which consists of the phosphorylation sites only shared between great apes and gibbons (<20~25 Mya). The frequency gradually decreases as the estimated time of origin increases, which reaches 34.0% for the 'oldest' group (150~100 Mya) and 29.6% for evolutionarily conserved group.

Evolution rates observed for phosphorylation sites also support the previous observation (Table 22). Average rate of evolution of S-P and T-P sites were 0.00263 (mutation / Mya) and 0.00292 respectively, while the rates for other serine and threonine sites were 0.00194 and 0.00225 respectively. P-values calculated with ANOVA indicates there is a significant difference between the rates of evolution of S/T-P sites and other S/T phosphorylation sites (P-value = 1.31E-03 for comparison between S-P and other S, P-value = 1.97E-02 for comparison between T-P and other T), while there was no significant difference between serine phosphorylation sites and threonine phosphorylation sites (P-value = 1.97E-01 for comparison between other S and other T, P-value = 2.13E-01 for comparison between S-P and T-P). The underlying reason is yet to be addressed, but strong association of S/T-P sites with IDRs & flexible regions,

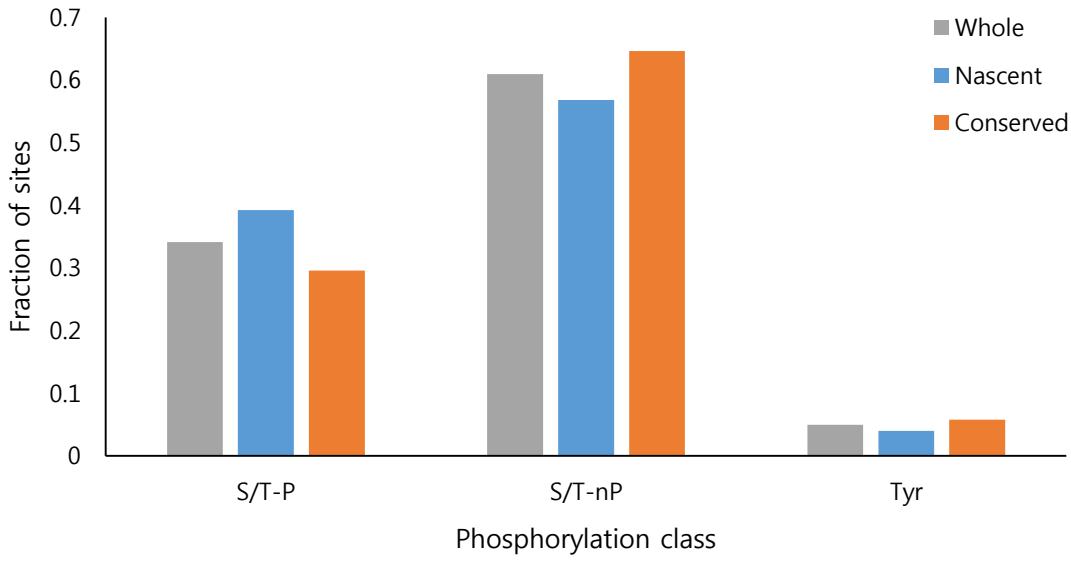

Figure 60. Composition of classes found in nascent / conserved phosphorylation sites

Class	Phos-site	Nonphos-site	ANOVA pair	p-value
S-nP	0.001943	0.005994	**S-nP/T-nP**	1.97E-1
T-nP	0.002249	0.002236	**S-P/T-P**	2.13E-1
Tyr	0.001758	0.001332		
S-P	0.002629	0.007089	**S-nP/S-P**	1.31E-3
T-P	0.002916	0.003116	**T-nP/T-P**	1.97E-2

Table 22. Rate of evolution observed in orthologs of S/T-P and S/T-nP sites

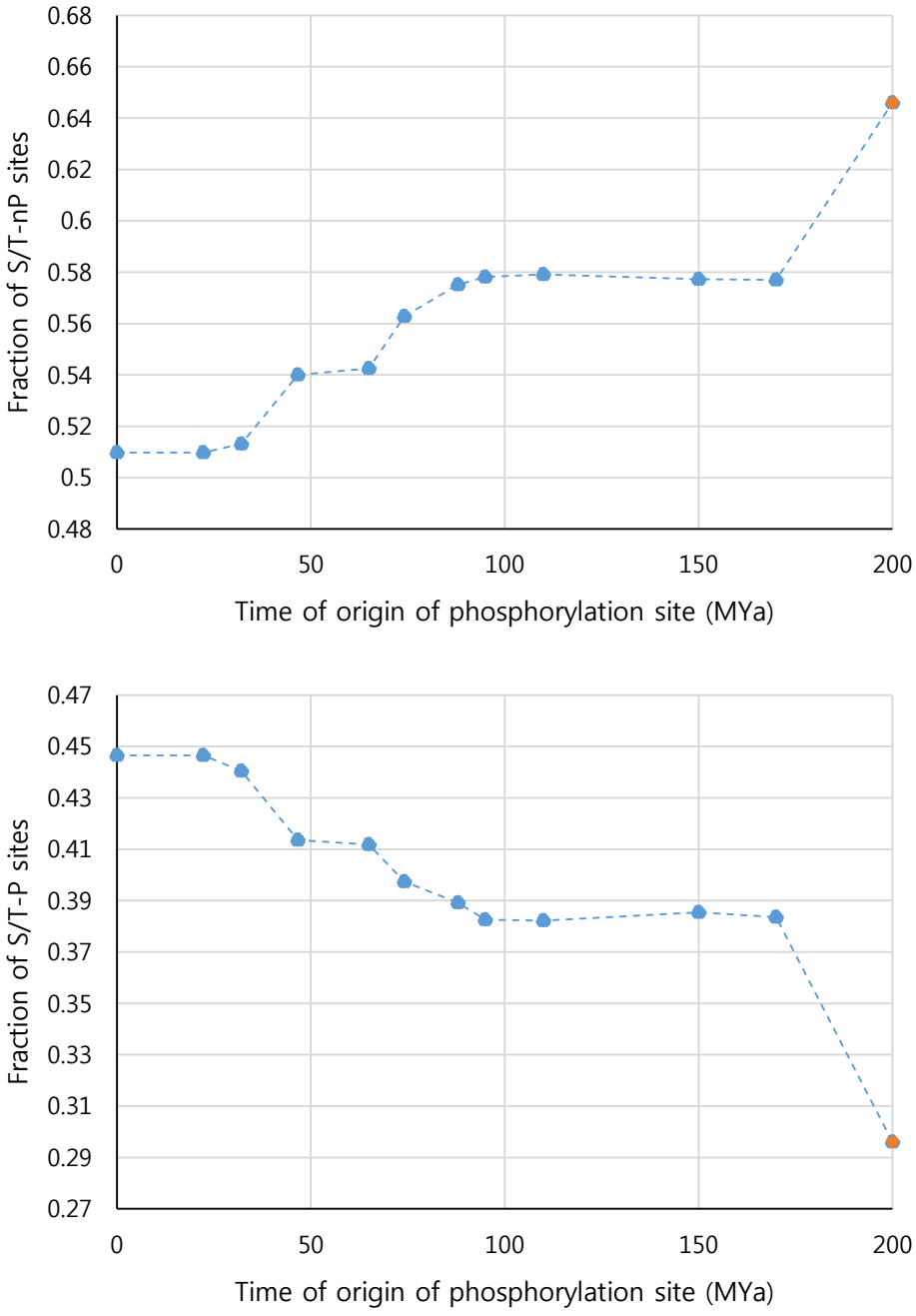

Figure 61. S/T-P site ratio is negatively correlated with estimated time-of-origin

which could tolerate high number of small mutations, could be one of possible explanations for this difference

-5.3.2. Higher occurrence rate of +1 proline is supported by the site-specific rates of mutation

Analysis of site-specific amino acid substitution rates revealed that proline residue could be accumulated at +1 site relative to phosphorylation site. Average proline to non-proline substitution rates observed in the alignment was 2.0E-03 (Figure 62). If the proline of interest is accompanied with non-phosphorylated serine or threonine, the rates were 1.1E-03 and 8.9E-04 respectively, showing a modest decrease. However, if the accompanying serine or threonine is known phosphorylation site, the observed rates were 3.3E-04 and 1.2E-04, which were 3.7-fold and 7.5-fold lower than values observed for non-phosphorylated SP / TP dipeptides. Comprehensive analysis of amino acid substitution rates showed +1 proline could be conserved better than any other amino acid at +1 site (Figure 63).

On the other hand, average non-proline to proline substitution rate was 6.2E-05 (Figure 62), much lower than that of the opposite event. The individual substitution values were 5.4E-05 for non-phosphorylated S-nP, 2.1E-05 for non-phosphorylated T-nP, 1.7E-04 for S-nP site and 1.6E-04 for T-nP site. Interestingly, substitution rate has increased when accompanying serine or threonine residues are known phosphorylation site: coupled with decreased proline to non-proline substitution rate, this could lead to the substantial enrichment of +1 proline observed in human phosphoproteome (Figure 62). Assuming that only +1 residue is modified, the equilibrium frequencies of +1 proline calculated from rates of substitution were 33.9% and 56.8% for phosphoserine and phosphothreonine respectively, which were strikingly similar to the observed frequencies of S-P sites (32.9%) and T-P sites (46.8%) in human proteome.

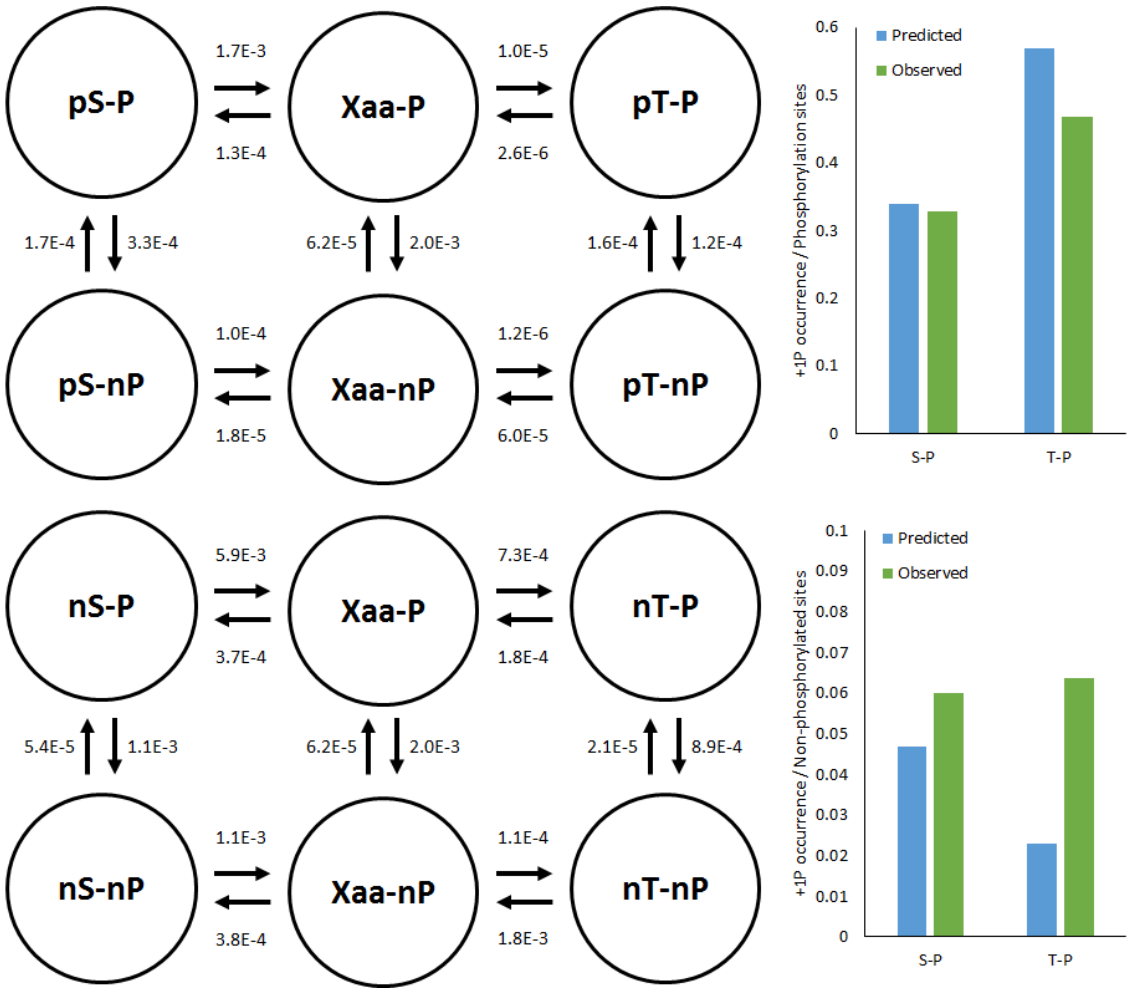

Figure 62. Rate of substitution between proline and non-proline amino acids (Upper left panel: rates of substitution observed in phosphorylation sites. Upper right panel: estimated equilibrium frequency of +1 proline on serine / threonine phosphorylation sites based on calculated substitution rates. Lower left panel: rates of substitution observed in non-phosphorylated sequences. Lower right panel: estimated equilibrium frequency of +1 proline on non-phosphorylated serine / threonine residues based on calculated substitution rates)

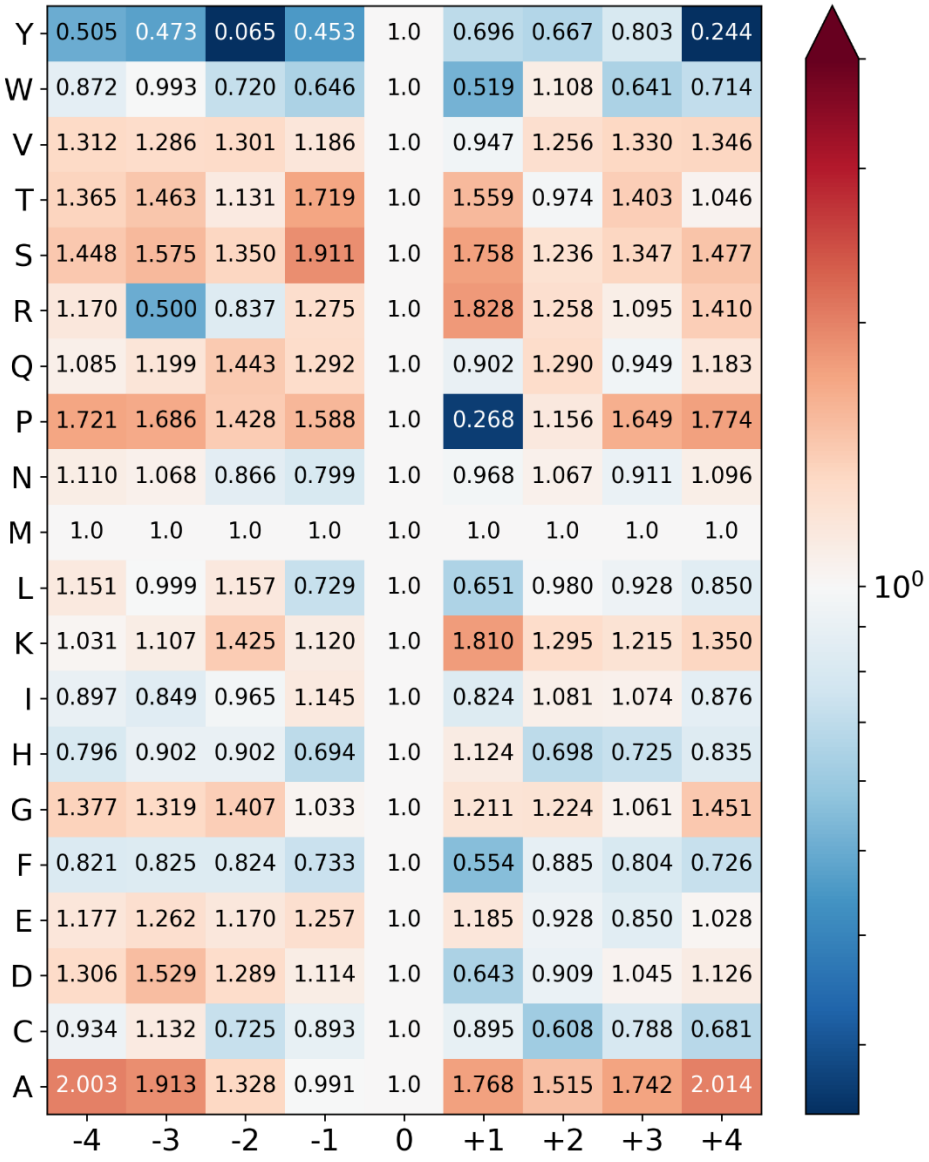

Figure 63. Relative rates of substitution from [Specific amino acid] to [Random amino acid] observed around phosphorylation site

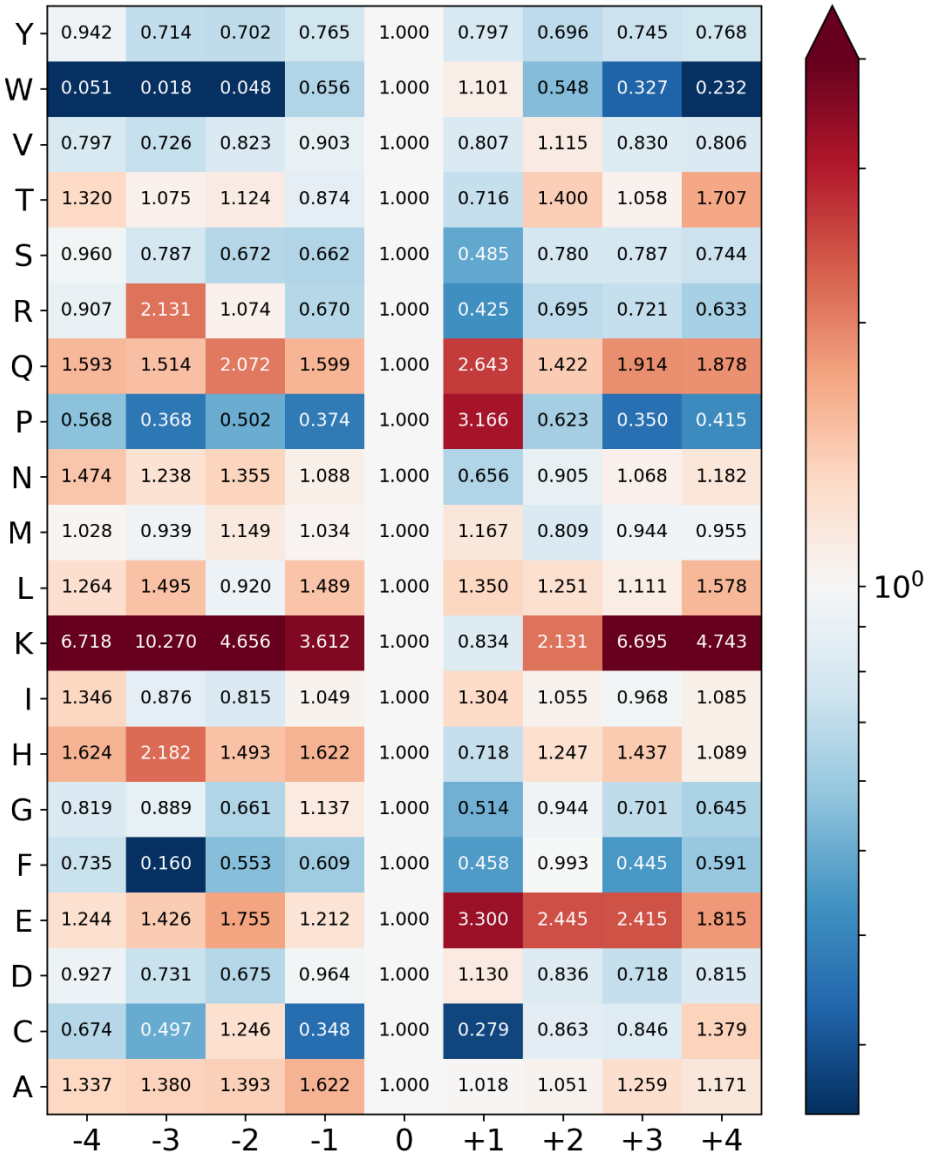

Figure 64. Relative rates of substitution from [Random amino acid] to [Specific amino acid] observed around phosphorylation site

	Phosphorylation site	Non-phosphorylated site	Relative rate of evolution
A	0.0027966	0.0015816	1.7682344
C	0.0047536	0.0053109	0.8950583
D	0.0005023	0.0007807	0.6433562
E	0.0015161	0.0012788	1.1855549
F	0.0015565	0.002807	0.5545192
G	0.0011784	0.0009724	1.2118559
H	0.0028182	0.002507	1.1241195
I	0.0014738	0.0017868	0.8248526
K	0.0025615	0.001415	1.8102108
L	0.0006897	0.0010586	0.6514817
M	0.0001001	0.0001001	1
N	0.0018832	0.0019454	0.9680018
P	0.0002887	0.001075	0.2685669
Q	0.0001224	0.0001356	0.902866
R	0.001204	0.0006584	1.8287706
S	0.0010755	0.0006116	1.758523
T	0.0018952	0.0012149	1.5599633
V	0.0011364	0.0011989	0.9478341
W	0.0021402	0.0041232	0.5190637
Y	6.60E-05	9.47E-05	0.696886

Table 23. Amino acid-specific evolution rates observed at +1 site and non-phosphorylated sequences

Calculation of two-sided substitution rates for all amino acids within ±4AA range (Figures 63, 64) revealed that other positive markers of phosphorylation sites, such as K/R at -3/-2 sites or glutamate at +2/+3 sites could be also enriched via evolutionary process. For instance, rates of substitution into lysine residues across the region except for +1 site were remarkable, but the rates of substitution from lysine residues to others were also increased – which resulted in slightly reduced enrichment. On the other hand, proline residues other than +1 proline showed clear sign of depletion, along with other aliphatic and aromatic residues.

-5.3.3. +1 proline is more likely to predate phosphorylated S/T residues in the ortholog alignment

Interesting discovery about S/T-P sites was that the +1 proline is more likely to be found in the ancestral sequence than the phosphorylation site itself. Reconstructed ancestral sequence for different levels of clades revealed that the number of ancestors which could not be phosphorylated at all is at least twofold higher than that of ancestors without +1 proline (Figure 65).

While low rate of substitution was observed for +1 proline than serine or threonine residues (Figure 62, 63), for now it is not clear whether it is a characteristic pattern of S/T-P phosphorylation sites or shared in all SP / TP dipeptide motifs. However, it was found that phosphorylated S/T and +1 proline were co-preserved (chi-square test resulted in p-value < 1.0E-4), making a probability of having an ancestral sequence with phosphorylated S/T but not +1 proline significantly lower than that expected from the null hypothesis (assumption of independence).

-5.3.4. Rate of substitution between serine and threonine at phosphorylation site was significantly different than the rate observed in other S/T residues

Another discovery should be mentioned is that the rate of substitution between serine and threonine (S↔T) residues were remarkably different, especially for S/T-nP sites. Due to the similar biochemistry and close codons, S↔T substitution is one of the most frequently found amino acid substitution in protein orthologs. Also, considering serine / threonine kinases could recognize both serine and threonine, S↔T substitution is

predicted not to affect the status as a phosphorylation site: as this substitution is 'neutral' in terms of functionality, it was anticipated that there would be no remarkable abnormalities regarding rate of S↔T substitution.

However, the observed rate of S↔T substitution questioned this presumed functional neutrality. Rates of substitution between non-phosphorylated serine and threonine residues were found to be almost the same regardless of accompanying +1 residue (Figure 66). The rate of S→T substitution was about 45~50-fold higher than that of T→S substitution process. In S/T-P sites, the fold difference between two rates of substitution increased to 67.75-fold, while its statistical significance is not clear enough to develop a hypothesis based on this observation. On the other hand, in S/T-nP sites, the fold difference was just 5.37-fold, with significantly higher rate of T→S substitution and decreased rate of substitution from S→T.

This could be interpreted as a mild preference towards phosphoserine in S/T-nP sites, which is supported by relatively lower frequency of threonine (Table 10) in S/T-nP sites (12.2%) than that found in S/T-P sites (20.6%) or non-phosphorylated sequences (38.9%). This might also reflect the hypothesized class-specific biases toward specific functionality (Figures 18-21) which would rely more on either site-specific interactions or changes in thermodynamic environments.

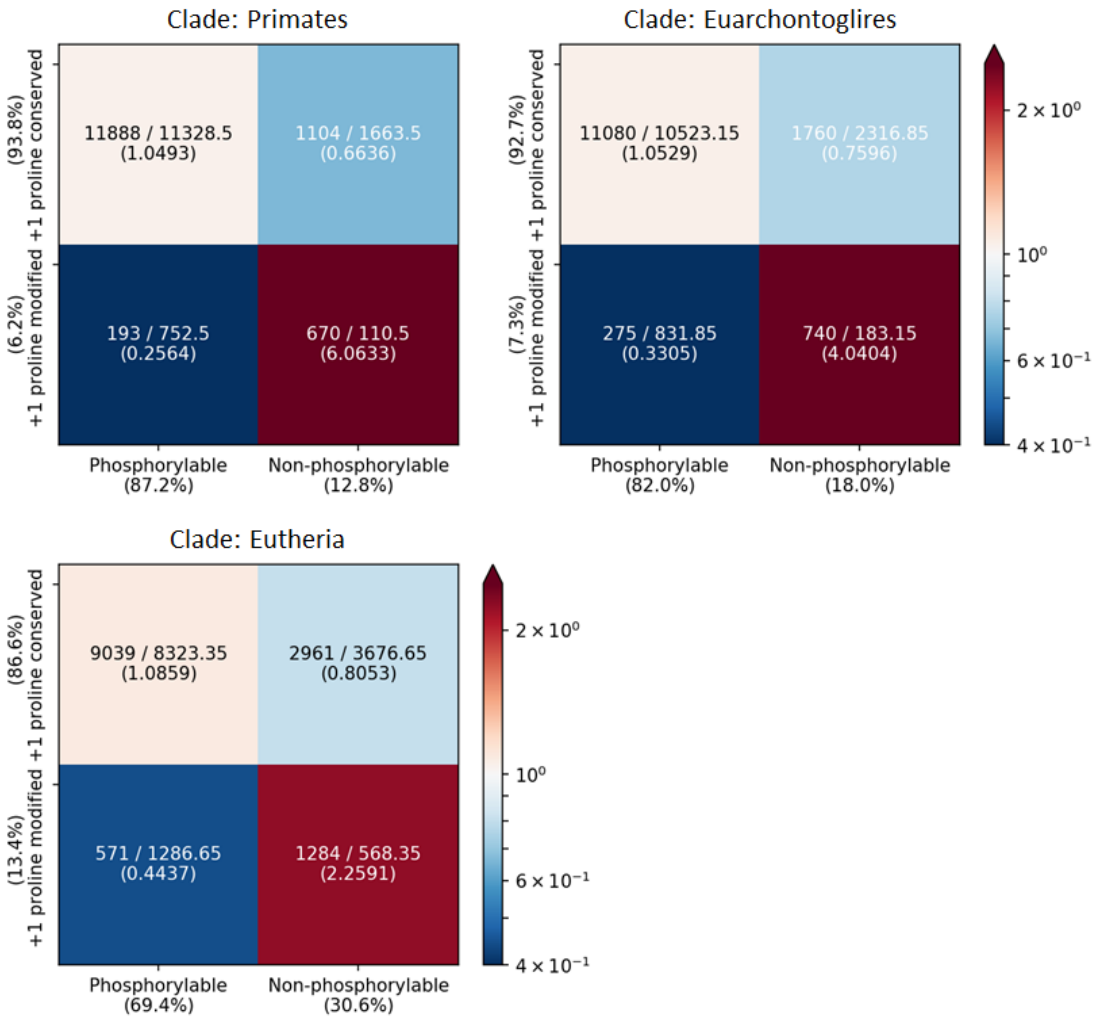

Figure 65. Conservation rates of phosphorylated S/T residues and +1 proline. Contingency tables were generated using ortholog sequences and reconstructed ancestral sequences for each clade. Colors of each cell denote fold change (observed / expected).

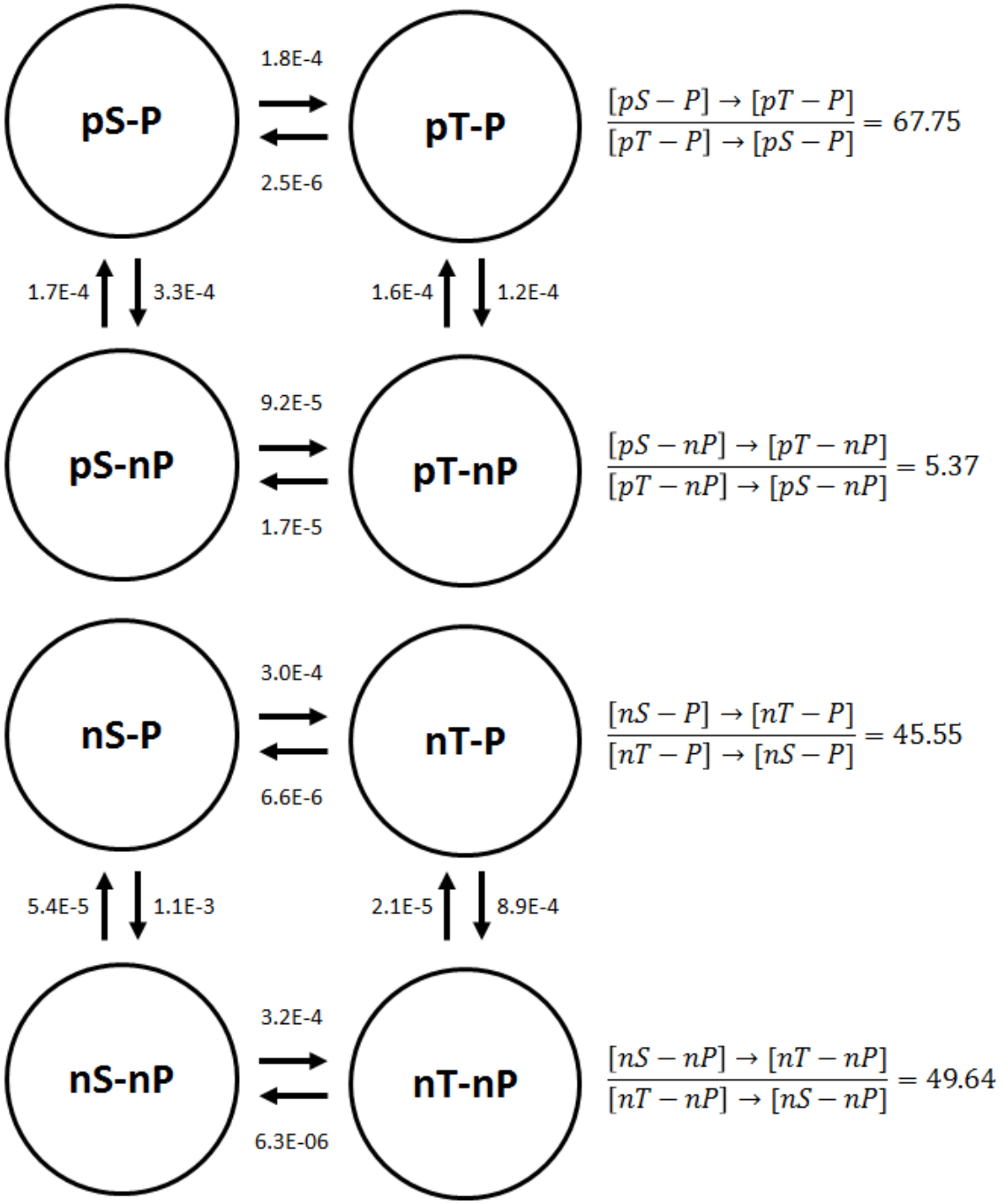

Figure 66. Rates of substitution between serine and threonine residues. (Upper panel: rates of substitution observed in phosphorylation sites. Lower panel: rates of substitution observed in non-phosphorylated sequences)

[5-4] Discussion

All these results suggest S/T-P sites are associated with evolutionary patterns distinguishable from those associated with other phosphorylation sites. This indicates S/T-P sites are associated with different evolutionary selection strategy, which might be influenced by aforementioned biophysical and biological differences.

S/T-P sites are more likely to have a younger origin than other types of phosphorylation sites. These nascent phosphorylation sites are significantly less likely to be accompanied with specific annotations about biological functions, raising a question whether the majority of S/T-P sites are non-functional phosphorylation sites which are just tolerated within IDRs or it is selected by a subtle evolutionary advantage which is hardly observable with traditional techniques. Clinical variant phenotypes (Figure 3) somewhat supports the first idea, as the modification of either phosphorylated S/T or +1 proline rarely resulted in pathogenic phenotypes. However, as the clinical data only provides large-scale phenotypes, it is highly possible that molecular-scale changes driven by phosphorylation were occluded by the activities of other components inside the cell, thereby rendering it practically not observable in tissue level and beyond. Also, enzymatic machineries specifically associated with S/T-P sites are often crucial for maintaining normal cell physiology, indicating S/T-P sites as a whole is indeed indispensable.

This leaves several questions; if the majority of S/T-P sites are at least minimally functional, how the individual S/T-P sites affect cell physiology?; how could we observe the effects of phosphorylation of S/T-P sites? Also, if the majority of S/T-P sites are actually non-functional, then how cells tolerate higher level of non-functional phosphorylation sites?; is there any reason why the CMGC kinases are so much promiscuous?; why the selection process allowed accumulation of S/T-P sites at the first place? I hope these questions to be properly addressed in the future, as the answers would provide valuable insights about how PTMs co-evolved with eukaryotic proteome and interactome.

Even though S/T-P sites seem to be relatively dynamically added and removed, distribution of S/T-P sites across the proteome was far from random (section 3.3.4), posing an interesting question. While there are multiple reports about biological properties of phosphorylation site pairs with a specific distance (187, 112,

188), no one is not sufficient to explain overall distribution of phosphorylation sites; in fact, lack of information about individual function of S/T-P sites blocks from associating sufficient number of phosphorylation sites with specific hypothesis, making statistical analysis implausible. Still, this result suggest there is a biological function associated with S/T-P sites which is substantial enough to generate negative selection pressure, thereby removing potentially deleterious phosphorylation sites from the protein sequence. Along with positive selection pressure associated with some phosphorylation patterns, such as +-1/2 double tyrosine phosphorylation (by autocatalysis) or +-4 double phosphorylation of substrates targeted by GSK3B, this negative selection pressure might be one of the mechanism which sculpts eukaryotic proteomes.

Another notable discovery was that the frequencies of substitution between serine and threonine were unexpectedly low. Even though it could happen with only a single base change, the substitution rates were actually quite different from the rates observed in non-phosphorylated S/T. This suggests that for the phosphorylation sites, serine and threonine is not completely synonymous and have different functionality at least partially. This hypothesis is supported by previous research dealing with conformational changes induced by phosphorylation (62, 107) and eSCAPE analysis results (section 3.3.2): in both cases, phosphothreonines have a greater effect on local secondary structure or thermodynamic descriptors than phosphoserines. However, there is no known example of S/T substitution inducing a tangible biological phenotype at least in the cellular level, and there is no known kinase which specifically prefers one type of amino acid over another. The only clue is that the frequency of threonines in whole S/T-P sites is significantly higher than that found in other S/T phosphorylation sites, which might be consistent with the proposed higher contribution of biophysical effects to the functions of S/T-P sites.

I would like to provide an examples of human-mouse ortholog pairs which the phosphorylation site is present in one species but absent in another species (Table 24). Unfortunately, none of the phosphorylation site was associated with specific biological function, but the proteins are relatively well characterized with biological activities. These pairs could be an accessible research opportunity to elucidate effects of S/T-P

phosphorylation sites on IDR stretches – the anecdotal evidence provided from the research would improve our understanding of S/T-P phosphorylation..

Table 24. Assorted phosphorylation sites which is only observed in one species (human / mouse) but not in another species

Phosphorylation site changed (S/T-P sites)

Human protein	Site	Human sequence	Mouse sequence
CAMK1	363	PGTEL**S**PTLPH	PGSEL**P**PAPPP
CUL7	339	QLADV**S**PGLPA	RPAQF**R**PYTQR
DAXX	668	ICTLP**S**PPSPL	TSVQP**M**PSPPL
GSK3B	390	QAAAS**T**PTNAT	QAAAS**P**PANAT
HTT	1870	STKLL**S**PQMSG	CTKSL**N**PQKSG
MDM1	83	SNVVA**S**PEPEA	KDTLV**P**PEPQA
PHF1	420	SVSPP**S**PSPNQ	SVSPP**P**PSPNQ
RAD9A	328	PSISL**S**PGPQP	PSTSL**P**PVSLA
RAD9A	380	SPQGP**S**PVLAE	SPQGP**N**PVLAE
UBR1	21	AELPQ**T**PQRLA	PEPPL**A**PQRPA

+1 residue changed (HUMAN S/T-P, MOUSE S/T-NP)

Human protein	Site	Human sequence	Mouse sequence
ACIN	365	EMKTTS**P**LEEE	ETQIVS**L**PQEE
AURKB	35	RKEPVT**P**SALV	RKEPAT**T**SALA
NEDD1	468	NVFMGS**P**GKEE	NVLMGS**S**GKEE
NEDD1	550	INGSST**P**NPKI	VNGSST**T**VPKA
RELB	37	LGALGS**P**DLSS	LGALGS**S**DLSS

+1 residue changed (HUMAN S/T-NP, MOUSE S/T-P)

Human protein	Site	Human sequence	Mouse sequence
ATP7B	481	AKSPQS**T**RAVA	SETPSS**P**GATA
CAMK4	12	PSCSAS**S**CSSV	PSCPSS**P**CSSV
PPIG	254	KKSKKS**A**SSES	KKSKKS**P**SSES
PRKDC	2612	VETQAS**Q**GTLQ	IETQAS**P**SILH
TP53BP1	830	PVEQDS**S**QPSL	AVTEDS**P**QPPL

[6] Discussion

Thorough investigation and characterization of phosphorylation sites revealed S/T-P phosphorylation sites are not only biophysically distinguishable but also associated with different biological functions and evolutionary history. This indicates that S/T-P sites should be considered as a genuine class of phosphorylation sites just as tyrosine phosphorylation sites does, which means the current paradigm of eukaryotic phosphorylation should be revised at least in some degree.

Why S/T-P phosphorylation sites become so common and widespread at the first place? I already mentioned CMGC kinases are numerous, have poor sequence specificity and recognize multiple phosphorylation sites as its substrate: this could be the easiest explanation for the stated question, but it only concerns about the current occurrence and ignores any evolutionary context. Therefore it could not be the fundamental reason why the S/T-P sites become enriched in complex eukaryotic proteomes.

On the other hand, eukaryotic proteome is characterized by high occurrence of serine and proline residues (Figure 67), which are all known to be disorder-promoting residues (156). This makes SP / TP dipeptides to be one of the most commonly found dipeptide in the protein sequence, especially in the IDR (Figure 68). Therefore, biological mechanism targeting SP / TP dipeptides could affect the largest number of IDRs while maintaining minimal sequence specificity. Possible examples include MAPKs and CDKs, which phosphorylate hundreds of downstream substrates and thereby induce a massive change of cell physiology.

Modulation of multiple sites with a single effector is indeed a common mode-of-regulation in eukaryotic cells: signaling cascades are the classic examples of this. There are other kinases which phosphorylate large number of substrates and significantly affect cellular environment (e.g. PKA, PKC, CAMKs), but the substrates targeted by these kinases have different biophysical properties (e.g. N'-terminal positive charge), different intracellular localization, and less strongly associated with IDRs overall. This suggests S/T-P phosphorylation shares the fundamental objective of regulating an increasing number of substrates simultaneously, but manifests differently to better fit with specific peptide environment - IDR.

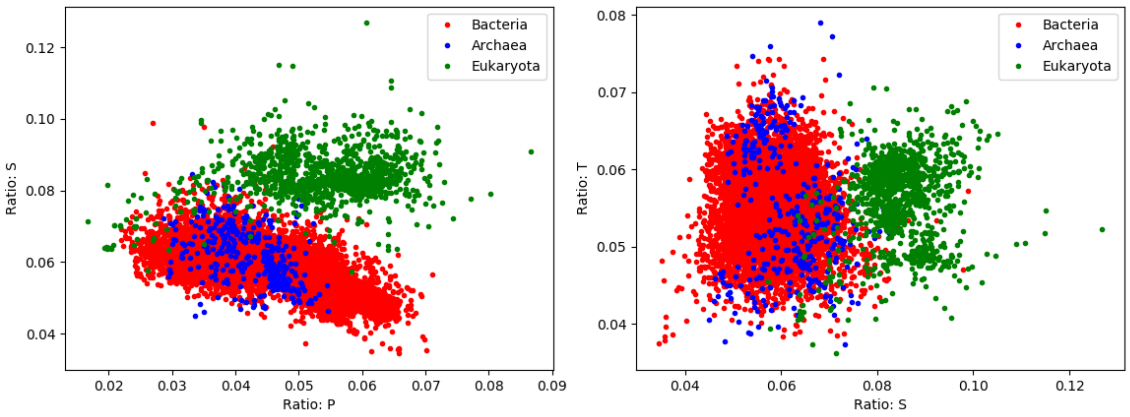

Figure 67. Frequency of proline / serine / threonine residues in reference proteomes

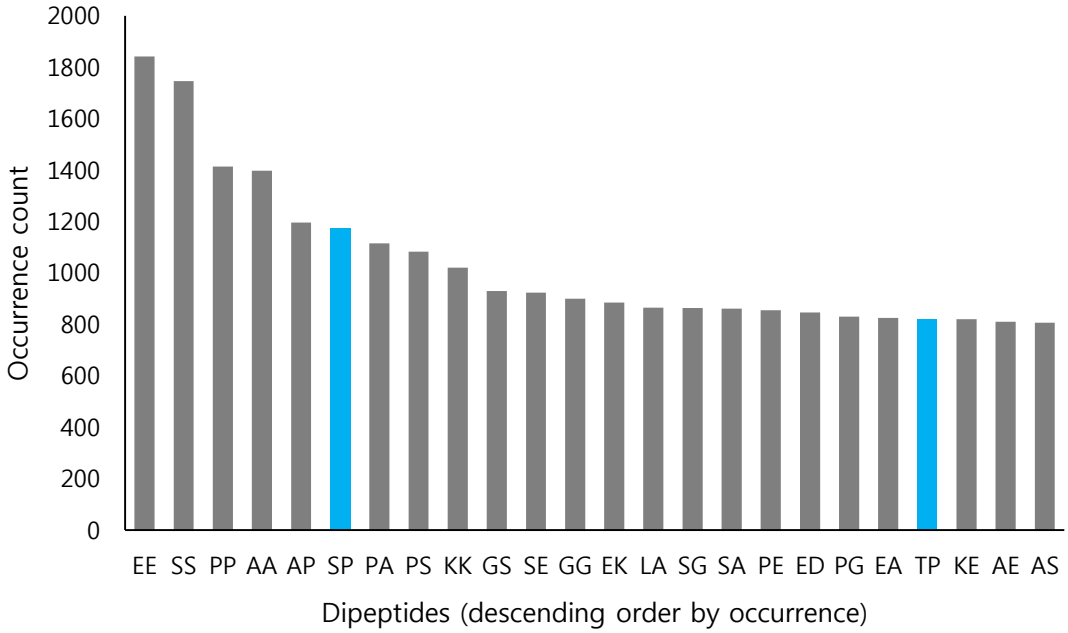

Figure 68. Dipeptide occurrences (top 25 dipeptides) in experimentally proven IDR

Even bolder hypothesis is that the S/T-P sites are the first example of 'dipeptide PTM' or 'indirect PTM' - a post translational modification which the modified residue is coupled with another residue which contributes more to the local dynamics ('biophysical effector'). Because of its aliphatic side chain, only few kinds of PTMs directly involve proline residue, and these PTMs have limited biophysical effects. On the other hand, phosphorylation have a profound biophysical effects which is strong enough to induce local conformational changes. Therefore, coupling S/T phosphorylation with +1 proline could be a mechanism of exploiting inherent properties of prolines in site-specific & reversible manner, which would have provided significant evolutionary advantages to eukaryotes - complex organism with high frequency of IDRs in particular.

This hypothesis is supported by the data which shows both phosphorylated S/T and +1 proline are co-preserved (Figure 65), and +1 proline is more likely to predate phosphorylated S/T, meaning properties of proline would be the essence of S/T-P site functionality. However, there is no known example which the modification of +1 proline causes observable phenotype changes, largely due to the combined effect of general lack of information about S/T-P sites and neglect of +1 proline as a functional compartment of PTM. Also, as proline itself is a special type of amino acid, identifying effects specifically associated with phosphorylation by missense mutation would be complicated, as mutation would also change the properties of non-phosphorylated peptide.

Therefore, finding direct evidence about this hypothesis would be a daunting task, but it would provide an important insight about how the paired phosphorylation site and proline generate an emergent property which affects thousands of different proteins. Detailed investigation of individual proteins with S/T-P sites, specifically those involved in signal transduction (189), liquid droplet/stress granule formation (190) or irreversible protein aggregation (17), could be one of the applicable options of this.

There are two other PTM-associated mechanisms which would interact with phosphorylation of S/T-P sites - O-glycosylation and hyperphosphorylation. O-glycosylation is a 'competitor PTM' - it targets S/T residue, and prefers +1 proline nearby modified site (46). While the annotation is relatively scarce, recent data suggests O-glycosylation would modify much higher number of substrates than we currently know, which increases the possibility of interference between two PTM types. However, O-glycosylation has a different

effect on local conformation: while phosphorylated substrates prefer alpha helix and polyproline II helix, O-glycosylated substrates prefer extended conformations akin to beta-sheet (191). This would allow cells to utilize two modes of PTMs as means of modulating IDRs and responding to different external signals more directly.

Hyperphosphorylation overrides effects of single phosphorylation of S/T-P sites with multiple charge-based effects, which possibly generates an additional conformational state. While S/T-P sites has a mild tendency of avoiding hyperphosphorylation clusters, the sheer number of S/T-P sites allow some of the site to be included in one of these clusters. There are two kinases noteworthy to mention - GSK3B and casein kinase 2 (CK2). GSK3B is a CMGC kinase which only phosphorylates substrates with already phosphorylated residues (156). CK2, which is previously considered as a CMGC kinases but reclassified as a different group recently, targets for substrates with negatively charged amino acids in C'-terminal side (192). These two kinases are indicative of the potential mechanism of sequential hyperphosphorylation triggered by phosphorylation of S/T-P sites, thereby providing cells another way to modulate IDRs.

[7] Concluding remarks

Until recently, S/T-P sites were neglected for a long time. There are actually some information could be connected to this: considering the nature of IDR regions, individual effect of S/T with P phosphorylation might be very subtle, hard to experimentally identify. This is justified by the nebulous data about S/T-P sites which become available so far: site-directed studies, clinical variants and experimental mutagenesis cultivated an indifference, consequently attracting less attention from the researchers. Even more, the data suggested these sites might not have a tangible functionality by itself: some of them may manifest its functionality when there are accompanying modifications or regulator molecules, but it is possible that the site might be a noise, no more than that.

However, the narrative of this work suggests otherwise: S/T-P sites have different sequence characteristics, separated enzyme subsets in both kinases and phosphatases, and even different evolutionary history. Implementation of conformational dynamics further elucidated the differences in biophysical properties which also affects the consequences of phosphorylation, which collectively demonstrates it is neither a biochemical noise nor byproducts of other serine/threonine phosphorylation. Recognizing this difference allowed our development of a state-of-the-art prediction algorithm which could be utilized for future research.

This work provides thorough evidence which indicates S/T-P phosphorylation sites are indeed a distinct subclass of PTM, which shares the phosphate group with classical phosphorylation sites but associated with different functionality. This particular PTM might be selected as a general regulation mechanism of IDRs, which has been continuously increasing after the emergence of multicellularity. IDRs, if not controlled, may cause significant problems inside the cell, spanning from premature degradation to non-specific interaction to irreversible aggregation. This became even more serious problem when the IDR contents increased along with the expansion of eukaryotic proteome, which exponentially increases the network complexity and threat of 'catastrophe'. The pre-mentioned virtues of phosphorylation makes S/T-P phosphorylation a great mechanism of general regulation: energy efficiency and high concentration of ATP allow multitude of substrates to be modified by a single signal immediately and reversibility allows those proteins to be salvaged if necessary. In this perspective, having no other sequence marker might even be a selected quality which

allows kinases to target even more substrates, which exist in lower concentration (in average). Alternatively, this could be understood as a modification of SP/TP dipeptide to trigger downstream effects dependent of properties of proline, which could not be modified as freely as other residues at the +1 site. If this point of view is indeed valid, this might be a first example of 'indirect PTM' or 'dipeptide PTM': which the modified residue and 'biophysical effector' residue are separated.

It would also provide a general idea about the ‘emergent property’ generated by coupling two distinct features common to all lifeforms – proline and protein phosphorylation in this case – which might have been exploited by eukaryotes to deal with unique challenge posed by increasing complexity and IDR contents. Further investigation of protein features and identification of similar emergent properties would not only allow a better understanding of protein biology but also provide another approach of enhancing practical applications, including rational protein design and drug discovery.

[8] References

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Analyzing Protein Structure and Function. In Molecular Biology of the Cell. 4th ed. New York: Garland Science; 2002.

2. Berg JM, Tymoczko JL, Stryer L. Protein Structure and Function. In Biochemistry. 5th edition. New York: W H Freeman; 2002.

3. Crick FH. Central dogma of molecular biology. Nature 1970; 227(5258):561-3.

4. Ambrogelly A, Palioura S, Söll D. Natural Expansion of the Genetic Code. Nat chem biol. 2007;3(1):29-35.

5. Walczak R, Westhof E, Carbon P, Krol A. A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. RNA. 1996;2(4):367-379.

6. Li WT, Mahapatra A, Longstaff DG, Bechtel J, Zhao G, Kang PT, Chan MK, Krzycki JA. Specificity of pyrrolysyl-tRNA synthetase for pyrrolysine and pyrrolysine analogs. J Mol Biol. 2009;385(4):1156–64.

7. Jensen ON. Modification-specific proteomics: characterization of post-translational modifications by mass spectrometry. Curr Opin Chem Biol. 2004;8:33-41.

8. Prabakaran S, Lippens G, Steen H, Gunawardena J. Post-translational modification: nature's escape from geneticimprisonment and the basisfor dynamic information encoding. Prabakaran, S., Lippens, G., Steen, H. & Gunawardena, J. Wiley Interdiscip. Rev Syst Biol Med. 20124:565–583.

9. Walsh CT, Garneau-Tsodikova S, Gatto GJ Jr. Protein posttranslational modifications: the chemistry of proteome diversifications. Angew Chem Int Ed Engl. 2005; 44(45):7342-72.

10. Theillet FX, Smet-Nocca C, Liokatis S, Thongwichian R, Kosten J, Yoon MK, Kriwacki RW, Landriew I, Lippens G, Selenko P. Cell signaling, post-translational protein modifications and NMR spectroscopy. J Biomol NMR. 2012;54:217–236.

11. Duan G, Walther D. The Roles of Post-translational Modifications in the Context of Protein Interaction Networks. PLoS Comput Biol. 2015;11(2):e1004049.

12. Karve TM, Cheema AK. Small changes huge impact: the role of protein posttranslational modifications in cellular homeostasis and disease. J Amino Acids. 2011;2011:207691.

13. Humphrey SJ, James DE, Mann M. Protein phosphorylation: a major switch mechanism for metabolic regulation. Trends Endocrinol Metab. 2015;26:676–68.

14. Karimi M, Ignasiak M, Chan B, Croft AK, Radom L, Schiesser CH, Pattison DI, Davies MJ. Reactivity of disulfide bonds is markedly affected by structure and environment: implications for protein modification and stability. Sci Rep. 2016;6:38572.

15. Komander D, Rape M. The ubiquitin code. Annu Rev Biochem. 2012;81:203–29.

16. Bowman GD, Poirier MG. Post-translational modifications of histones that influence nucleosome dynamics. Chem Rev. 2015;115:2274-2295.

17. Šimić G et al. Tau Protein Hyperphosphorylation and Aggregation in Alzheimer's Disease and Other Tauopathies, and Possible Neuroprotective Strategies. Biomolecules. 2016;6(1):6.

18. Huang K, Lee T, Kao H, Ma C, Lee C, Lin T, Chang W, Huang H. dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. Nuc Acids Res. 2019;47(D1):D298-D308.

19. Aebersold R. et al. How many human proteomforms are there? Nat Chem Biol. 14(3):206-214; 2018

20. Ezkurdia I. et al. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. Hum Mol Genet. 2014;23:5866-78.

21. Raju TS. Phosphorylation of proteins. In Co- and Post-Translational Modifications of Therapeutic Antibodies and Proteins. John Wiley & Sons Inc. 2019.

22. Potel CM, Lin M, Heck AJR, Lemeer S. Widespread bacterial protein histidine phosphorylation revealed by mass spectrometry-based proteomics. Nat Methods. 2018;15:187-190.

23. Pearlman SM, Serber Z, Ferrell JE Jr. A mechanism for the evolution of phosphorylation sites. Cell. 2011;147(4):934–946.

24. Mann M, Ong SE, Grønborg M, Steen H, Jensen ON, Pandey A. Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. Trends Biotechnol. 2002;20(6):261-8.

25. The uniprot consortium. UniProt: a worldwide hub of protein knowledge. Nuc Acids Res. 2019;47(D1):D506-515.

26. Park JO, Rubin SA, Xu YF, Amador-Noguez D, Fan J, Shlomi T, Rabinowitz JD. Metabolite concentrations, fluxes and free energies imply efficient enzyme usage. Nat Chem Biol. 2016;12(7):482–489.

27. Knorre DG, Kudryashova NV, Godovikova TS. Chemical and functional aspects of posttranslational modification of proteins. Acta Naturae. 2009;1(3):29–51.

28. Xie NZ, Du QS, Li JX, Huang RB. Exploring Strong Interactions in Proteins with Quantum Chemistry and Examples of Their Applications in Drug Design. Plos ONE. 2015;10(9):e0137113.

29. Sindelar CV, Hendsch ZS, Tidor B. Effects of salt bridges on protein structure and design. Protein sci. 1988;7:1898-1914.

30. Ardito F, Giuliani M, Perrone D, Troiano G, Lo Muzio L. The crucial role of protein phosphorylation in cell signaling and its use as targeted therapy (Review). Int. J. Mol. Med. 2017;40:271–280.

31. Turnham RE, Scott JD. Protein kinase A catalytic subunit isoform PRKACA; History, function and physiology. Gene. 2016;577(2):101–108.

32. Takemoto-Kimura S, Suzuki K, Horigane SI, Kamijo S, Inoue M, Sakamoto M, Fujii H, Bito H. Calmodulin kinases: essential regulators in health and disease. J Neurochem. 2017;141:808-818

33. Rasmussen H, Takuwa Y, Park S. Protein kinase C in the regulation of smooth muscle contraction. FASEB J. 1987;1(3):177-85.

34. Schittek B, Sinnberg T. Biological functions of casein kinase 1 isoforms and putative roles in tumorigenesis. Mol Cancer. 2014;13:231.

35. Hornbeck PV, Zhang B, Murray B, Kornhauser JM, Latham V, and Skrzypek E. PhosphoSitePlus, 2014:

mutations, PTMs and recalibrations. Nuc Acids Res. 43:D512-20; 2015.

36. Amanchy R, Kandasamy K, Mathivanan S, et al. Identification of Novel Phosphorylation Motifs Through an Integrative Computational and Experimental Analysis of the Human Phosphoproteome. J Proteomics Bioinform. 2011;4(2):22–35.

37. Bórquez DA, González-Billault C. Bioinformatics Approaches for Predicting Kinase–Substrate Relationships. In Bioinformatics - Updated Features and Applications. IntechOpen. 2016

38. Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol. 1999;294(5):1351-62.

39. Plotnikov A, Zehorai E, Procaccia S, Seger R. The MAPK cascades: signaling components, nuclear roles and mechanisms of nuclear translocation. Biochim Biophys Acta. 2011 Sep;1813(9):1619-33.

40. Chen W, Dang T, Blind RD, Wang Z, Cavasotto CN, Hittelman AB, Rogatsky I, Logan SK, Garabedian MJ. Glucocorticoid receptor phosphorylation differentially affects target gene expression. Mol Endocrinol. 2008 Aug;22(8):1754-66.

41. Fry AM, O'Regan L, Sabir SR, Bayliss R. Cell cycle regulation by the NEK family of protein kinases. J Cell Sci. 2012;125(Pt 19):4423–4433.

42. Lundby A, Secher A, Lage K, et al. Quantitative maps of protein phosphorylation sites across 14 different rat organs and tissues. Nat Commun. 2012;3:876.

43. Guo Y, Yang K, Harwalkar J, Nye JM, Mason DR, Garrett MD, Hitomi M, Stacey DW. Phosphorylation of cyclin D1 at Thr 286 during S phase leads to its proteasomal degradation and allows efficient DNA synthesis. Oncogene. 2005 Apr 14;24(16):2599-612.

44. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nuc Acids Res. 2014;D42:D980–D985.

45. Koscielny G, Yaikhom G, Iyer V, et al. The International Mouse Phenotyping Consortium Web Portal, a unified point of access for knockout mice and related phenotyping data. Nuc Acids Res. 2014;D42:D802-D809.

46. Theillet FX, Kalmar L, Tompa P, et al. The alphabet of intrinsic disorder: I. Act like a Pro: On the abundance and roles of proline residues in intrinsically disordered proteins. Intrinsically Disord Proteins. 2013;1(1):e24360.

47. Vacic V, Uversky VN, Dunker AK, Lonardi S. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. BMC Bioinformatics. 2007;8:211.

48. Berger A, Kurtz J, Katchalski E. Poly-L-proline. J Am Chem Soc. 1954;76:5552-4.

49. MacArthur MW, Thornton JM. Influence of proline residues on protein conformation. J Mol Biol. 1991;218(2):397-412.

50. Viguera AR, Serrano L. Stable proline box motif at the N-terminal end of alpha-helices. Protein Sci. 1999;8(9):1733–1742.

51. Pal D, Chakrabarti P. Cis peptide bonds in proteins: residues involved, their conformations, interactions and locations. J Mol Biol. 1999;294(1):271-88.

52. Steinberg IZ, Harrington WF, Berger A, Sela M, Katchalski E. The configurational changes of poly-L-proline in solution. J Am Chem Soc. 1960;82:5263-79.

53. Tomasso ME, Tarver MJ, Devarajan D, Whitten ST. Hydrodynamic Radii of Intrinsically Disordered Proteins Determined from Experimental Polyproline II Propensities. PLoS Comput Biol. 2016;12(1):e1004686.

54. Krieger F, Möglich A, Kiefhaber T. Effect of proline and glycine residues on dynamics and barriers of loop formation in polypeptide chains. J Am Chem Soc. 2005;127(10):3346-52.

55. Chiu CC, Singh S, de Pablo JJ. Effect of Proline Mutations on the Monomer Conformations of Amylin. Biophys J. 2013;105(5):1227-1235

56. Rauscher S, Baud S, Miao M, Keeley FW, Pomès R. Proline and glycine control protein self-organization into elastomeric or amyloid fibrils. Structure. 2006;14(11):1667-76.

57. Gorres KL, Raines RT. Prolyl 4-hydroxylase. Crit Rev Biochem Mol Biol. 2010;45(2):106–124.

58. DeRider ML, Wilkens SJ, Waddell MJ, Bretscher LE, Weinhold F, Raines RT, Markley JL. Collagen stability: Insights from NMR spectroscopic and hybrid density functional computational investigations of the effect of electronegative substituents on prolyl ring conformations. J Am Chem Soc. 2002;124:2497–2505.

59. Shoulders MD, Raines RT. Collagen structure and stability. Annu Rev Biochem. 2009;78:929–958.

60. Hon WC1, Wilson MI, Harlos K, Claridge TD, Schofield CJ, Pugh CW, Maxwell PH, Ratcliffe PJ, Stuart DI, Jones EY. Structural basis for the recognition of hydroxyproline in HIF-1 alpha by pVHL. Nature. 2002;417(6892):975-8.

61. Wilson IB, Gavel Y, von Heijne G. Amino acid distributions around O-linked glycosylation sites. Biochem J. 1991;275:529–534.

62. Elbaum MB, Zondlo NJ. OGlcNAcylation and Phosphorylation Have Similar Structural Effects in α-Helices: Post-Translational Modifications as Inducible Start and Stop Signals in α-Helices, with Greater Structural Effects on Threonine Modification. Biochemistry. 2014;53(14):2242-2260

63. Brister MA, Pandey AK, Bielska AA, Zondlo NJ. OGlcNAcylation and phosphorylation have opposing structural effects in tau: phosphothreonine induces particular conformational order. J Am Chem Soc. 2014;136(10):3803–3816.

64. Rani L, Mallajosyula SS. Phosphorylation versus O-GlcNAcylation: Computational Insights into the Differential Influences of the Two Competitive Post-Translational Modifications. J Phys Chem B. 2017;121(47):10618-10638

65. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science. 2002;298(5600):1912-34.

66. Varjosalo M, Keskitalo S, Van Drogen A, Nurkkala H, Vichalkovski A, Aebersold R, Gstaiger M. The protein interaction landscape of the human CMGC kinase group. Cell Rep. 2013;3(4):1306-20.

67. Malumbres M. Cyclin-dependent kinases. Genome Biol. 2014;15(6):122.

68. Becker W. Emerging role of DYRK family protein kinases as regulators of protein stability in cell cycle control. Cell Cycle. 2012;11(18):3389–3394.

69. Aubol BE, Chakrabarti S, Ngo J, et al. Processive phosphorylation of alternative splicing factor/splicing factor 2. Proc Natl Acad Sci U S A. 2003;100(22):12601–12606.

70. Kondo S, Lu Y, Debbas M, et al. Characterization of cells and gene-targeted mice deficient for the p53-binding kinase homeodomain-interacting protein kinase 1 (HIPK1). Proc Natl Acad Sci U S A. 2003;100(9):5431–5436.

71. Brown NR, Noble ME, Endicott JA, Johnson LN. The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. Nat Cell Biol. 1999;1(7):438-43.

72. Kummer L, Parizek P, Rube P, et al. Structural and functional analysis of phosphorylation-specific binders of the kinase ERK from designed ankyrin repeat protein libraries. Proc Natl Acad Sci U S A. 2012;109(34):E2248–E2257.

73. Kannan N, Neuwald AF. Evolutionary constraints associated with functional specificity of the CMGC protein kinases MAPK, CDK, GSK, SRPK, DYRK, and CK2alpha. Protein Sci. 2004;13(8):2059–2077.

74. Howard CJ, Hanson-Smith V, Kennedy KJ, et al. Ancestral resurrection reveals evolutionary mechanisms of kinase plasticity. Elife. 2014;3:e04126.

75. Obenauer JC, Cantley LC, Yaffe MB. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nuc Acids Res. 2003;31(13):3635–3641.

76. Wedemeyer WJ, Welker E, Scheraga HA. Proline cis-trans isomerization and protein folding. Biochemistry. 2002;41(50):14637-44.

77. Schmidpeter PA, Schmid FX. Prolyl isomerization and its catalysis in protein folding and protein function. J Mol Biol. 2015;427(7):1609-31.

78. Wang CC, Tsou CL. Enzymes as chaperones and chaperones as enzymes. FEBS Lett. 1998; 425:382-4.

79. Chen Y, Wu YR, Yang HY, et al. Prolyl isomerase Pin1: a promoter of cancer and a target for therapy. Cell Death Dis. 2018;9(9):883.

80. Liou YC, Zhou XZ, Lu KP. Prolyl isomerase Pin1 as a molecular switch to determine the fate of phosphoproteins. Trends Biochem Sci. 2011;36(10):501–514.

81. Lu Z, Hunter T. Prolyl isomerase Pin1 in cancer. Cell Res. 2014;24(9):1033–1049.

82. Ryo A. et al. Prolyl-isomerase Pin1 accumulates in lewy bodies of parkinson disease and facilitates formation of alpha-synuclein inclusions. J Biol Chem. 2006;281(7):4117-25.

83. Balastik M, Lim J, Pastorino L, Lu KP. Pin1 in Alzheimer's disease: multiple substrates, one regulatory mechanism?. Biochim Biophys Acta. 2007;1772(4):422–429.

84. Kimura T, Tsutsumi K, Taoka M, et al. Isomerase Pin1 stimulates dephosphorylation of tau protein at cyclin-dependent kinase (Cdk5)-dependent Alzheimer phosphorylation sites. J Biol Chem. 2013;288(11):7968–7977.

85. Innes BT, Bailey ML, Brandl CJ, Shilton BH, Litchfield DW. Non-catalytic participation of the Pin1 peptidyl-prolyl isomerase domain in target binding. Front Physiol. 2013;4:18.

86. Babu MM. The contribution of intrinsically disordered regions to protein function, cellular complexity,

and human disease. Biochem Soc Trans. 2016;44(5):1185–1200.

87. Fersht AR. From the first protein structures to our current knowledge of protein folding: delights and scepticisms. Nat Rev Mol Cell Biol. 2008;9(8):650-4.

88. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol. 2007;8(12):995-1005.

89. Uversky VN. Intrinsically Disordered Proteins and Their "Mysterious" (Meta)Physics. Front Phys. 2019;7:10.

90. Oates ME, Romero P, Ishida T, et al. D²P²: database of disordered protein predictions. Nucleic Acids Res. 2013;D41:D508–D516.

91. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. Intrinsic disorder and protein function. Biochemistry. 2002;41(21):6573-82.

92. van der Lee R, Buljan M, Lang B, et al. Classification of intrinsically disordered regions and proteins. Chem Rev. 2014;114(13):6589–6631.

93. Hilser VJ, Thompson EB. Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. Proc Natl Acad Sci U S A. 2007;104(20):8311–8315.

94. Haynes C, Oldfield CJ, Ji F, et al. Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol. 2006;2(8):e100.

95. Zhao Y, Garcia BA. Comprehensive Catalog of Currently Documented Histone Modifications. Cold Spring Harb Perspect Biol. 2015;7(9):a025064.

96. Migliaccio AR, Uversky VN. Dissecting physical structure of calreticulin, an intrinsically disordered Ca2+-buffering chaperone from endoplasmic reticulum. J Biomol Struct Dyn. 2018;36(6):1617–1636.

97. Motlagh HN, Anderson JA, Li J, Hilser VJ. Disordered allostery: lessons from glucocorticoid receptor. Biophys Rev. 2015;7(2):257–265.

98. Yoon MK, Mitrea DM, Ou L, Kriwacki RW. Cell cycle regulation by the intrinsically disordered proteins p21 and p27. Biochem Soc Trans. 2012;40(5):981-8.

99. Dennis MK, Taneva SG, Cornell RB. The intrinsically disordered nuclear localization signal and phosphorylation segments distinguish the membrane affinity of two cytidylyltransferase isoforms. J Biol Chem. 2011;286(14):12349–12360.

100. Darling AL, Uversky VN. Intrinsic Disorder and Posttranslational Modifications: The Darker Side of the Biological Dark Matter. Front Genet. 2018;9:158.

101. Pejaver V, Hsu WL, Xin F, Dunker AK, Uversky VN, Radivojac P. The structural and functional signatures of proteins that undergo multiple events of post-translational modification. Protein Sci. 2014;23(8):1077–1093.

102. Iakoucheva LM, Radivojac P, Brown CJ, et al. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 2004;32(3):1037–1049.

103. Wright PE, Dyson HJ. Intrinsically disordered proteins in cellular signalling and regulation. Nat Rev

Mol Cell Biol. 2015;16(1):18–29.

104. Flock T, Weatheritt RJ, Latysheva NS, Babu MM. Controlling entropy to tune the functions of intrinsically disordered regions. Curr Opin Struct Biol. 2014;26:62-72.

105. Chin AF, Toptygin D, Elam WA, Schrank TP, Hilser VJ. Phosphorylation Increases Persistence Length and End-to-End Distance of a Segment of Tau Protein. Biophys J. 2016;110(2):362–371.

106. Groban ES, Narayanan A, Jacobson MP. Conformational changes in protein loops and helices induced by post-translational phosphorylation. PLoS Comput Biol. 2006;2(4):e32.

107. Elam WA, Schrank TP, Campagnolo AJ, Hilser VJ. Evolutionary conservation of the polyproline II conformation surrounding intrinsically disordered phosphorylation sites. Protein Sci. 2013;22(4):405–417.

108. Monahan Z, Ryan VH, Janke AM, et al. Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. EMBO J. 2017;36(20):2951–2967.

109. Kumar A, Gopalswamy M, Wolf A, Brockwell DJ, Hatzfeld M, Balbach J. Phosphorylation-induced unfolding regulates p19INK4d during the human cell cycle. Proc Natl Acad Sci U S A. 2018;115(13):3344–3349.

110. So L, Lee J, Palafox M, et al. The 4E-BP-eIF4E axis promotes rapamycin-sensitive growth and proliferation in lymphocytes. Sci Signal. 2016;9(430):ra57.

111. Sharma A, Singh K, Almasan A. Histone H2AX phosphorylation: a marker for DNA damage. Methods Mol Biol. 2012;920:613-26.

112. Beurel E, Grieco SF, Jope RS. Glycogen synthase kinase-3 (GSK3): regulation, actions, and diseases. Pharmacol Ther. 2015;148:114–131.

113. Amoutzias GD, Bornberg-Bauer E, Oliver SG, Robertson DL. Reduction/oxidation-phosphorylation control of DNA binding in the bZIP dimerization network. BMC Genomics. 2006;7:107.

114. Hegde RS, Kang SW. The concept of translocational regulation. J Cell Biol. 2008;182(2):225–232.

115. Hunter T. Tyrosine phosphorylation: thirty years and counting. Curr Opin Cell Biol. 2009;21(2):140–146.

116. Stancik IA, Šestak MS, Ji B, Axelson-Fisk M, Franjevic D, Jers C, Domazet-Lošo T, Mijakovic I. Serine/Threonine Protein Kinases from Bacteria, Archaea and Eukarya Share a Common Evolutionary Origin Deeply Rooted in the Tree of Life. J Mol Biol. 2018;430(1):27-32.

117. Dinkel H, Chica C, Via A, et al. Phospho.ELM: a database of phosphorylation sites--update 2011. Nucleic Acids Res. 2011;D39:D261–D267.

118. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics. 2015;15(18):3163–3168.

119. Thomsen MC, Nielsen M. Seq2Logo: a method for construction and visualization of amino acid binding motifs and sequence profiles including sequence weighting, pseudo counts and two-sided representation of amino acid enrichment and depletion. Nucleic Acids Res. 2012;W40:W281–W287.

120. Ashburner et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25-9.

121. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019;47(D1):D330-D338.

122. Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. Nucleic Acids Res. 2019;47(D1):D419-D426.

123. Miller CJ, Turk BE. Homing in: Mechanisms of Substrate Targeting by Protein Kinases. Trends Biochem Sci. 2018;43(5):380–394.

124. Walte A, Rüben K, Birner-Gruenberger R, Preisinger C, Bamberg-Lemper S, Hilz N, Bracher F, Becker W. Mechanism of dual specificity kinase activity of DYRK1A. FEBS J. 2013;280(18):4495-511.

125. Macossay-Castillo M, Marvelli G, Guharoy M, Jain A, Kihara D, Tompa P, Wodak SJ The balancing act of intrinsically disordered proteins: enabling functional diversity while minimizing promiscuity. J. Mol. Biol. 2019;431(8):1650-1670

126. Miranda-Saavedra D, Barton GJ. Classification and functional annotation of eukaryotic protein kinases. Proteins. 2007;68(4):893-914.

127. Shankar A, Agrawal N, Sharma M, Pandey A, Girdhar K Pandey M. Role of Protein Tyrosine Phosphatases in Plants. Curr Genomics. 2015;16(4):224–236.

128. Taylor SS, Radzio-Andzelm E, Hunter T. How do protein kinases discriminate between serine/threonine and tyrosine? Structural insights from the insulin receptor protein-tyrosine kinase. FASEB J. 1995;9(13):1255-66.

129. Tu Z, Young A, Murphy C, Liang JF. The pH sensitivity of histidine-containing lytic peptides. J Pept Sci. 2009;15(11):790–795.

130. Zhao N, Pang B, Shyu CR, Korkin D. Charged residues at protein interaction interfaces: unexpected conservation and orchestrated divergence. Protein Sci. 2011;20(7):1275–1284.

131. Anjana R, Vaishnavi MK, Sherlin D, et al. Aromatic-aromatic interactions in structures of proteins and protein-DNA complexes: a study based on orientation and distance. Bioinformation. 2012;8(24):1220–1224.

132. Karimi M, Ignasiak MT, Chan B, et al. Reactivity of disulfide bonds is markedly affected by structure and environment: implications for protein modification and stability. Sci Rep. 2016;6:38572.

133. Liao SM, Du QS, Meng JZ, Pang ZW, Huang RB. The multiple roles of histidine in protein interactions. Chem Cent J. 2013;7(1):44.

134. Schauperl M, Podewitz M, Waldner BJ, Liedl KR. Enthalpic and Entropic Contributions to Hydrophobicity. J Chem Theory Comput. 2016;12(9):4600–4610.

135. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157(1):105-32.

136. Hilser VJ, Freire E. Structure-based calculation of the equilibrium folding pathway of proteins: Correlation with hydrogen exchange protection factors. J Mol Biol. 1996;262(5):756-72.

137. Pang CN, Hayen A, Wilkins MR. Surface accessibility of protein post-translational modifications. J Proteome Res. 2007;6(5):1833-45.

138. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963;7:95-9.

139. Nordlund TM, Hoffman PM. Structures: from 0.1 to 10nm and larger. In Quantitative Understanding of Biosystems: An Introduction to Biophysics. CRC Press. 2011.

140. Jiang Q, Jin X, Lee SJ, Yao S. Protein secondary structure prediction: A survey of the state of the art. J Mol Graph Model. 2017;76:379-402.

141. Mansiaux Y, Joseph AP, Gelly JC, de Brevern AG. Assignment of PolyProline II conformation and analysis of sequence--structure relationship. PLoS One. 2011;6(3):e18401.

142. Li SS. Specificity and versatility of SH3 and other proline-recognition domains: structural basis and implications for cellular signal transduction. Biochem J. 2005;390(Pt 3):641–653.

143. Mikhonin AV, Myshakina NS, Bykov SV, Asher SA. UV resonance Raman determination of polyproline II, extended 2.5(1)-helix, and beta-sheet Psi angle energy landscape in poly-L-lysine and poly-L-glutamic acid. J Am Chem Soc. 2005;127(21):7712-20.

144. Tomasso ME, Tarver MJ, Devarajan D, Whitten ST. Hydrodynamic Radii of Intrinsically Disordered Proteins Determined from Experimental Polyproline II Propensities. PLoS Comput Biol. 2016;12(1):e1004686.

145. Vertrees J, Barritt P, Whitten S, Hilser VJ. COREX/BEST server: a web browser-based program that calculates regional stability variations within protein structures. Bioinformatics. 2005;21(15):3318-9.

146. Gu J, Hilser VJ. Predicting the energetics of conformational fluctuations in proteins from sequence: a strategy for profiling the proteome. Structure. 2008;16(11):1627–1637.

147. Mao AH, Crick SL, Vitalis A, Chicoine CL, Pappu RV. Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. Proc Natl Acad Sci U S A. 2010;107(18):8183–8188.

148. Cho MH, Wrabl JO, Hilser VJ, Taylor J. Hidden dynamic signatures drive substrate selectivity in the disordered phosphoproteome. Proc Natl Acad Sci U S A. 2020 (Under revision)

149. Robertson NO, Smith NC, Manakas A, et al. Disparate binding kinetics by an intrinsically disordered domain enables temporal regulation of transcriptional complex formation. Proc Natl Acad Sci U S A. 2018;115(18):4643–4648.

150. Altenhoff AM, Glover NM, Train CM, et al. The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucleic Acids Res. 2018;46(D1):D477–D485.

151. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A. 1992;89(22):10915–10919.

152. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M. AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. 2008;D36:D202–D205.

153. Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK. TOP-IDP-scale: a new amino

acid scale measuring propensity for intrinsic disorder. Protein Pept Lett. 2008;15(9):956–963.

154. Hatos A, Hajdu-Soltész B, Monzon AM, et al. DisProt: intrinsic protein disorder annotation in 2020. Nucleic Acids Res. 2020;48(D1):D269-D276

155. Das RK, Pappu RV. Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proc Natl Acad Sci U S A. 2013;110(33):13392–13397.

156. Uversky VN. The alphabet of intrinsic disorder: II. Various roles of glutamic acid in ordered and intrinsically disordered proteins. Intrinsically Disord Proteins. 2013;1(1):e24684.

157. Schad E, Tompa P, Hegyi H. The relationship between proteome size, structural disorder and organism complexity. Genome Biol. 2011;12(12):R120.

158. Xin F, Radivojac P. Post-translational modifications induce significant yet not extreme changes to protein structure. Bioinformatics. 2012;28(22):2905-13.

159. Andrew CD, Warwicker J, Jones GR, Doig AJ. Effect of phosphorylation on alpha-helix stability as a function of position. Biochemistry. 2002;41(6):1897-905.

160. Hanes SD. The Ess1 prolyl isomerase: traffic cop of the RNA polymerase II transcription cycle. Biochim Biophys Acta. 2014;1839(4).

161. Needham EJ, Parker BL, Burykin T, James DE, Humphrey SJ. Illuminating the dark phosphoproteome. Sci Signal. 2019;12(565).

162. Que S, Wang Y, Chen P, Tang YR, Zhang Z, He H. Evaluation of protein phosphorylation site predictors. Protein Pept Lett. 2010;17(1):64-9.

163. Loganantharaj R. Extensions of Naive Bayes and Their Applications to Bioinformatics. In Bioinformatics Research and Applications. ISBRA 2007. Lecture Notes in Computer Science, vol 4463. Springer, Berlin, Heidelberg. 2007.

164. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot. 2013;7:21.

165. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. Ann Stat. 2001;29(5):1189-1232.

166. Wrabl JO, Larson SA, Hilser VJ. Thermodynamic propensities of amino acids in the native state ensemble: implications for fold recognition. Protein Sci. 2001;10(5):1032–1045.

167. Liu T, Pantazatos D, Li S, Hamuro Y, Hilser VJ, Woods VL Jr. Quantitative assessment of protein structural models by comparison of H/D exchange MS data with exchange behavior accurately predicted by DXCOREX. J Am Soc Mass Spectrom. 2012;23(1):43–56.

168. Hoffmann J, Wrabl JO, Hilser VJ. The role of negative selection in protein evolution revealed through the energetics of the native state ensemble [published correction appears in Proteins. 2018 Dec;86(12):1313]. Proteins. 2016;84(4):435–447.

169. Wang S, Gu J, Larson SA, Whitten ST, Hilser VJ. Denatured-state energy landscapes of a protein structural database reveal the energetic determinants of a framework model for folding. J Mol Biol. 2008;381(5):1184–1201.

170. Gao J, Thelen JJ, Dunker AK, Xu D. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. Mol Cell Proteomics. 2010;9(12):2586–2600.

171. Dou Y, Yao B, Zhang C. PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. Amino Acids. 2014;46(6):1459-69.

172. Wei L, Xing P, Tang J, Zou Q. PhosPred-RF: A Novel Sequence-Based Predictor for Phosphorylation Sites Using Sequential Information Only. IEEE Trans Nanobioscience. 2017;16(4):240-247.

173. Ismail HD, Jones A, Kim JH, Newman RH, Kc DB. RF-Phos: A Novel General Phosphorylation Site Prediction Tool Based on Random Forest. Biomed Res Int. 2016;2016:3281590.

174. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science. 1999;286(5438):295-9.

175. Schaefer C, Schlessinger A, Rost B. Protein secondary structure appears to be robust under in silico evolution while protein disorder appears not to be. Bioinformatics. 2010;26(5):625–631.

176. Trost B, Kusalik A, Napper S. Computational Analysis of the Predicted Evolutionary Conservation of Human Phosphorylation Sites. PLoS One. 2016;11(4):e0152809.

177. Boekhorst J, van Breukelen B, Heck A Jr, Snel B. Comparative phosphoproteomics reveals evolutionary and functional conservation of phosphorylation across eukaryotes. Genome Biol. 2008;9(10):R144.

178. Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. Curr Opin Struct Biol. 2009;19(5):596-604.

179. Huerta-Cepas J, Szklarczyk D, Heller D, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47(D1):D309–D314.

180. Larkin MA. et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23(21):2947-8.

181. Joy JB, Liang RH, McCloskey RM, Nguyen T, Poon AF. Ancestral Reconstruction. PLoS Comput Biol. 2016;12(7):e1004763.

182. Athey J, Alexaki A, Osipova E, et al. A new and updated resource for codon usage tables. BMC Bioinformatics. 2017;18(1):391.

183. Schneider A, Cannarozzi GM, Gonnet GH. Empirical codon substitution matrix. BMC Bioinformatics. 2005;6:134.

184. Bevan RB, Lang BF, Bryant D. Calculating the evolutionary rates of different genes: a fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. Syst Biol. 2005 ;54(6):900-15.

185. Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. Placental mammal diversification and the Cretaceous–Tertiary boundary. Proc Natl Acad Sci U S A. 2003;100(3):1056-1061

186. dos Reis M, Yang Z. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol Biol Evol. 2011;28(7):2161-72.

187. Valk E, Venta R, Ord M, Faustova I, Kõivomägi M, Loog M. Multistep phosphorylation systems: tunable components of biological signaling circuits. Mol Biol Cell. 2014;25(22):3456–3460.

188. Li Y, Zhou X, Zhai Z, Li T. Co-occurring protein phosphorylation are functionally associated. PLoS Comput Biol. 2017;13(5):e1005502.

189. Lu KP, Liou YC, Zhou XZ. Pinning down proline-directed phosphorylation signaling. Trends Cell Biol. 2002;12(4):164-72.

190. Larson AG, Elnatan D, Keenen MM, et al. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. Nature. 2017;547(7662):236–240.

191. Brister MA, Pandey AK, Bielska AA, Zondlo NJ. OGlcNAcylation and phosphorylation have opposing structural effects in tau: phosphothreonine induces particular conformational order. J Am Chem Soc. 2014;136(10):3803–3816.

192. St-Denis N, Gabriel M, Turowec JP, Gloor GB, Li SS, Gingras AC, Litchfield DW. Systematic investigation of hierarchical phosphorylation by protein kinase CK2. J Proteomics. 2015 Apr 6;118:49-62.

[9] Curriculum vitae

Min-Hyung Cho

Contact information

500 W. University Parkway, Apt 2F

Baltimore, MD, 21210 U.S.A.

(443) 676-9186

E-mails: mcho22@jhu.edu

sompi@naver.com

Personal information

Citizenship: Republic of Korea

Languages: Korean, English and Japanese

Education

- Ph.D. Biology. Johns Hopkins University (JHU) U.S.A. 2020 (expected).

  (M.S. Biological science. Seoul National University (SNU). Seoul – Republic of Korea. Did not finished: attended 2013 – 2014)

- B.S. Biological science. Seoul National University (SNU). Seoul – Republic of Korea, 2013.

Dissertations and Publications

- Ph.D. thesis: Reclassification of serine / threonine phosphorylation sites with +1 proline (S/T-P) sites as a distinct eukaryotic post-translational modification class

- Article: Min-Hyung Cho, James O. Wrabl, James Taylor and Vincent J. Hilser. Hidden dynamic signatures substrate selectivity in the disordered phosphoproteome. PNAS. 2020 (Under revision)

- Article: Min-Hyung Cho, Jong-Woo Kim and Guhung Jung, 2-amino-N-(2,6-dichloropyridin-3-yl)acetamide derivatives as a novel class of HBV capsid assembly inhibitor. J Viral Hepat. 2013.

- Article (bachelor project): Min-Hyung Cho, Jin-su Song, Hie-Joon Kim, Sung-Gyoo Park and Guhung Jung. Structure-based design and biological evaluation of sulfanilamide derivatives as hepatitis B virus capsid assembly. J Enzyme Inhib Med Chem. 2012.

Presentations and Proceedings

- Min Hyung Cho, Vincent Hilser & James Taylor. Thermodynamic profiles of phosphoproteins suggest a general fundamental role for serine/threonine phosphorylation sites with +1 proline (S/T-P) in eukaryotes. Poster session presented at the Biophysical Society Meeting 2020. San Diego.

- Min Hyung Cho, James O. Wrabl, Vincent Hilser & James Taylor. Phosphorylation sites with S/T-P motif: possible basal anti-aggregation mechanism. Poster session presented at the Biophysical Society Meeting 2019. Baltimore.
- Min Hyung Cho, James O. Wrabl, Vincent Hilser & James Taylor. PHOSforUS: Biophysical property-based protein phosphorylation predictor. Poster session presented at the Intelligent systems for molecular biology (ISMB) conference 2018. Chicago.
- Min Hyung Cho and Guhung Jung. The 'third physiological state' derived by prolonged exposure to ROS and its relationship to HBV chronic infection. Poster session presented at the SFRBM 2013 annual meeting.

Awards & Scholarships

- Victor Corces Teaching Award. JHU. 2016.
- Doctoral study abroad program (5 year scholarship). Korea Foundation for Advanced Studies (KFAS). 2014-2019

Teaching experience

- Teaching assistant of 250 undergraduate students in cell biology class. JHU. Biology department. 2019
- Teaching assistant of 20 undergraduate students in cell biology laboratory class. JHU. Biology department. 2016
- Teaching assistant of 20 undergraduate students in biochemistry laboratory class. JHU. Biology department. 2015
- Teaching assistant of 150 undergraduate students in genetics class. SNU. Department of Biological science. 2013.

Computer skills

- Operative system: Linux and Windows
- Programming: C, R, Python and Perl
- Biophysics: Origin and Pymol
- Bioinformatics: Bioconductor, Biopython, Galaxy
- Math software: Mathematica and Maple
- Office suit: MS office, Openoffice and LaTeX